Periodic Regression in Biology and Climatology

C. I. Bliss

Bulletin 615

June 1958

THE CONNECTICUT AGRICULTURAL EXPERIMENT STATION • NEW HAVEN

Foreword

Periodic phenomona in biology and climatology occur so widely that we tend either to adapt to them as unavoidable nuisances or are overimpressed by their day to day deviations. We can't "see the forest for the trees." If the variable occurs around the clock or through the year, but with systematically unequal magnitudes, its underlying pattern can often be expressed logically in relatively simple trigonometric terms. When this classic mathematical model is combined with an appropriate statistical analysis, we are better able both to describe the periodic trend and to study deviations from its pattern. For example we can separate weather into its orderly and its random elements and by this means estimate the probability of occurrence of critical temperatures. This approach is sufficiently novel, even to biologists and climatologists with a background in modern statistics, that the technique is described here in some detail. Its applications are illustrated with a wide range of biological examples and a more detailed study of a typical climatological series.

Periodic Regression in Biology and Climatology

C. I. Bliss

Most non-linear regressions in biology and many in climatology are handled in one of two ways. The first is to convert the relation to a straight line by the selection, on either theoretical or empirical grounds, of a suitable unit for each variable, such as its reciprocal, logarithm, probit or logit. A second approach is to fit a polynomial equation relating the dependent variable y to successive functions of the independent variable x. In one familiar form, these functions are the powers of x, leading to an equation of the form

$$Y = a + b_1x + b_2x^2 + b_3x^3 + \ldots + b_kx^k \qquad (1)$$

Given $k + 1$ values of our independent variable, the curve defined by this equation will fit exactly the mean responses $\bar{y}_i$ at each x, if extended to k powers of x. In practice, we terminate the series as soon as the residual variation of $\bar{y}_i$ about the fitted curve is comparable with the variation of the individual y's about their respective means.

When the relation between x and y is periodic, our polynominal equation will be more rational if we substitute trigonometric functions of x for their powers, leading to harmonic or Fourier analysis, or "periodic regression" as it is termed by Aitken (1939). The problem is further simplified when the independent variable x is cyclical in character with a length fixed independently of the response. Typical variables include the hour of day in the diurnal cycle, the month or week in the annual cycle, and the compass direction in dispersion from a center. We are not concerned here with cycles determined *a posteriori,* such as from fluctuations in the abundance of animals or of plant pests, nor with "cycles" which represent an age trend in a single group of individuals, such as the monthly egg production from

a single set of pullets through the year. We will further assume that each of the equally-spaced subdivisions in the cycle is represented by a constant number of observations. Within these restrictions, periodic regression parallels the more familiar curvilinear regression in which the orthogonal polynomials represent the successive powers of x.

The Sine Curve

Many periodic biological functions can be fitted by the symmetrical sine curve. We start with f values of our dependent variable y at each of k observed times t (or other interval) within the cycle. The expected response Y at each t may then be computed from the sine curve, expressed conveniently in the form

$$Y = a_o + A \cos(ct - \theta) \qquad (2)$$

where $a_o = \bar{y}$ is the mean response over f complete periods or cycles. The coefficient A is the semi-amplitude or one-half the range from the maximum to the minimum Y. The constant $c = 2\pi/k$ converts the numbered units of time, $t = 0, 1, 2, \ldots, k\text{-}1$, in a single cycle to angular measure in radians. The statistic θ is the phase angle or the time in angular measure of the maximum response Y. It shifts the origin for measuring time from an arbitrary starting point t_o to the time at which the response is a maximum. The angles could be measured equally in degrees instead of in

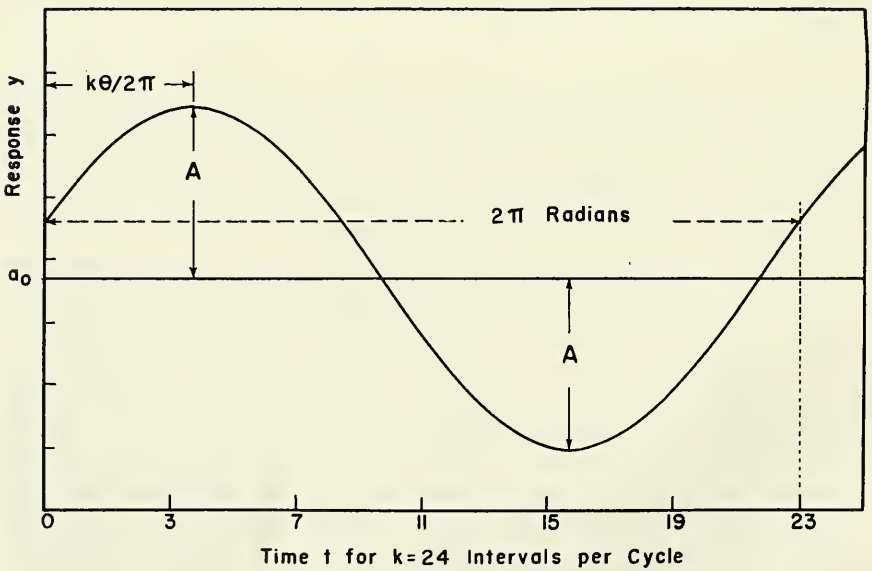

Figure 1. The sine curve and its constants.

radians, but radians have been selected here as the more convenient. One complete cycle of $360° = 2\pi = 6.283185$ radians. These various functions of the sine curve are shown graphically in Figure 1.

For estimating its constants from the observed responses, we may rewrite Equation 2 as

$$Y = a_o + a_1\cos(ct) + b_1\sin(ct) \tag{3}$$

an equation linear in the adjustable parameters a_1 and b_1, where

$$A = \sqrt{a_1^2 + b_1^2} \tag{4}$$

and

$$\tan\theta = b_1/a_1 \tag{5}$$

The expected response Y for a given t can be computed directly from Equation 3 without conversion to the original form. The range in units of y is equal to twice the semi-amplitude or 2A. To determine the correct

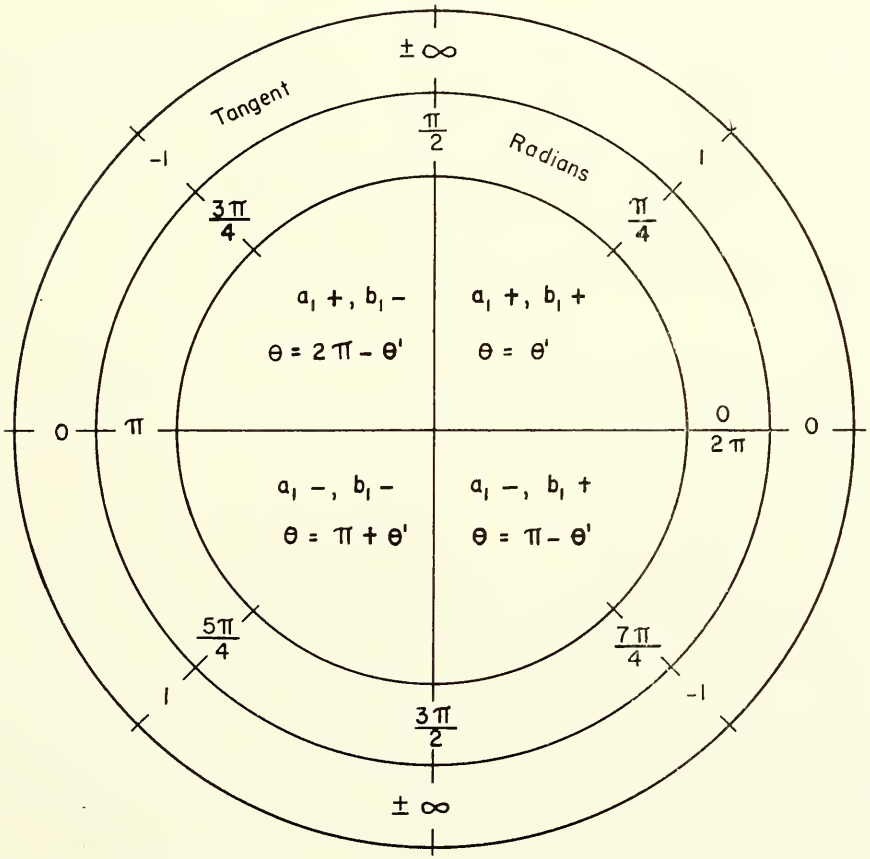

Figure 2. Conversion of $\theta' = |b_1/a_1|$ to the phase angie θ.

TABLE 1. Cosines (u_i) and sines (v_i) for the harmonic analysis of cyclical data recorded in k equally-spaced fractions per cycle and numbered consecutively from t = 0 to t = k−1.

k = 7

t	u_1	v_1	u_2	v_2	u_3	v_3
0	1	0	1	0	1	0
1	.6235	.7818	−.2225	.9749	−.9010	.4339
2	−.2225	.9749	−.9010	−.4339	.6235	−.7818
3	−.9010	.4339	.6235	−.7818	−.2225	.9749
4	−.9010	−.4339	.6235	.7818	−.2225	−.9749
5	−.2225	−.9749	−.9010	.4339	.6235	.7818
6	.6235	−.7818	−.2225	−.9749	−.9010	−.4339

k = 8

t	u_1	v_1	u_2	v_2	u_3	v_3	u_4
0	1	0	1	0	1	0	1
1	.707	.707	0	1	−.707	.707	−1
2	0	1	−1	0	0	−1	1
3	−.707	.707	0	−1	.707	.707	−1
4	−1	0	1	0	−1	0	1
5	−.707	−.707	0	1	.707	−.707	−1
6	0	−1	−1	0	0	1	1
7	.707	−.707	0	−1	−.707	−.707	−1

k = 12

t	u_1	v_1	u_2	v_2	u_3	v_3	u_4	v_4
0	1	0	1	0	1	0	1	0
1	.866	.5	.5	.866	0	1	−.5	.866
2	.5	.866	−.5	.866	−1	0	−.5	−.866
3	0	1	−1	0	0	−1	1	0
4	−.5	.866	−.5	−.866	1	0	−.5	.866
5	−.866	.5	.5	−.866	0	1	−.5	−.866
6	−1	0	1	0	−1	0	1	0
7	−.866	−.5	.5	.866	0	−1	−.5	.866
8	−.5	−.866	−.5	.866	1	0	−.5	−.866
9	0	−1	−1	0	0	1	1	0
10	.5	−.866	−.5	−.866	−1	0	−.5	.866
11	.866	−.5	.5	−.866	0	−1	−.5	−.866

k = 24

t	u_1	v_1	u_2	v_2
0	1	0	1	0
1	.966	.259	.866	.5
2	.866	.5	.5	.866
3	.707	.707	0	1
4	.5	.866	−.5	.866
5	.259	.966	−.866	.5
6	0	1	−1	0
7	−.259	.966	−.866	−.5
8	−.5	.866	−.5	−.866
9	−.707	.707	0	−1
10	−.866	.5	.5	−.866
11	−.966	.259	.866	−.5
12	−1	0	1	0
13	−.966	−.259	.866	.5
14	−.866	−.5	.5	.866
15	−.707	−.707	0	1
16	−.5	−.866	−.5	.866
17	−.259	−.966	−.866	.5
18	0	−1	−1	0
19	.259	−.966	−.866	−.5
20	.5	−.866	−.5	−.866
21	.707	−.707	0	−1
22	.866	−.5	.5	−.866
23	.966	−.259	.866	−.5

For t = 0-11 and 12-23:
$u_2,v_2 = u_1,v_1$ (k=12)
$u_4,v_4 = u_2,v_2$ (k=12)

For t = 0-7, 8-15, 16-23:
$u_3,v_3 = u_1,v_1$ (k=8)

For k = 4: $u_1,v_1 = u_2,v_2$ (k=8, t=0-3)

For k = 6: $u_1,v_1 = u_2,v_2$ (k=12, t=0-5; $u_2,v_2 = u_1,v_1$ (k=12, t=0-5)

quadrant for the phase angle θ, we first determine from a table of trigonometric functions the angle in radians corresponding to $\tan \theta' = |b_1/a_1|$, and from the signs of the coefficients a_1 and b_1 convert θ' to the phase angle θ by Figure 2 (Brooks and Carruthers, 1953). Then on the time scale measured from t_o, the maximum response occurs at the time $k\theta/2\pi$. Since the sine curve is symmetrical, the time for the minimum is one-half cycle before or after the time of the maximum.

For any selected series of k equally-spaced intervals in each complete cycle, the cosines and sines corresponding to the successive intervals of $t = 0, 1, 2, \ldots k\text{-}1$ are listed in the columns for u_1 and v_1 in Table 1. Each forms an orthogonal set of independent variates (within a negligible rounding error) similar to the orthogonal polynomials for the successive powers of x. With $u_1 = \cos(ct)$ and $v_1 = \sin(ct)$, Equation 3 may be written as

$$Y = a_o + a_1u_1 + b_1v_1 \tag{6}$$

where $\Sigma u_1 = \Sigma v_1 = \Sigma(u_1v_1) = 0$. The cosines and sines in Table 1 cover the series encountered most commonly and include the higher harmonics required for the Fourier analysis in the next section. Except for rounding errors, which usually may be neglected, the denominator of a_1 and of b_1 is the same for all evenly-spaced series of the same length k, or $\Sigma u_1^2 = \Sigma v_1^2 \doteq \frac{1}{2}k$. With this short-cut, the regression coefficients for a single measure at each time t ($f = 1$) are readily computed as

$$\begin{aligned} a_1 &= \Sigma(u_1y)/\Sigma u_1^2 = [u_1y]/\tfrac{1}{2}k \\ \text{and} \qquad b_1 &= \Sigma(v_1y)/\Sigma v_1^2 = [v_1y]/\tfrac{1}{2}k \end{aligned} \tag{7}$$

With f replicated y's at each t, totalling T_t, the regression coefficients are computed directly from the T_t's as

$$\begin{aligned} a_1 &= \Sigma(u_1T_t)/f\Sigma u_1^2 = [u_1T_t]/\tfrac{1}{2}fk \\ \text{and} \qquad b_1 &= \Sigma(v_1T_t)/f\Sigma v_1^2 = [v_1T_t]/\tfrac{1}{2}fk \end{aligned} \tag{8}$$

As an example of simple periodic regression, we may fit a sine curve to the monthly mean temperatures in New Haven (Table 2), for the 14 years from July 1943, when the Weather Bureau station was moved to its present location at the municipal airport, through June 1957. The totals T_t in the last row of Table 2 were multiplied by the variates u_1 and v_1 in Table 1 for $k = 12$ to obtain by Equation 8 the regression coefficients $a_1 = 1763.0944/84 = 20.9892$ and $b_1 = 292.7604/84 = 3.4852$. With these coefficients and the mean, $a_o = 8528.6/168 = 50.7655$, the expected Y for each month has been computed by Equation 6 and the corresponding variates u_1 and v_1 in Table 1. The Y's have been plotted as the curve in Figure 3, together with the observed monthly means $\bar{y}_t$. In this as in most

other figures the first few months have been repeated at the end, so as to emphasize the cyclic character of the relation. Inspection indicates a good fit; how good we will test more fully in a later section.

From these records the seasonal range or amplitude in the mean temperature at New Haven is $2A = 2\sqrt{20.9892^2 + 3.4852^2} = 42.553°F$ as estimated from the sine curve by Equation 4. To determine the time of the maximum (Equation 5), we may compute $\tan \theta' = 3.4852/20.9892 = 0.16605$ and from a trigonometric table, interpolate the angle θ' corresponding to this tangent. With both a_1 and b_1 positive, θ falls in the first quadrant (Figure 2), so that $\theta = \theta' = 0.16455$ radians and the maximum temperature is reached at $12\ \theta/2\pi = 1.9746/6.2832 = 0.3143$ months from our starting point (t_0) in the annual cycle. Since t_0 corresponds to mid-July, this places the maximum temperature in New Haven approximately at July 25 over these 14 years and the minimum six months later on January 24. These estimates, of course, are subject to sampling errors which will be considered in a later section. Apart from their intrinsic interest, they permit rewriting the prediction equation in Equation 6 in the form of Equation 2, if this is preferred, as

$$Y = 50.765° + 21.2766 \cos(0.5236t - 0.16455)$$

where t is the number of the month (Table 1).

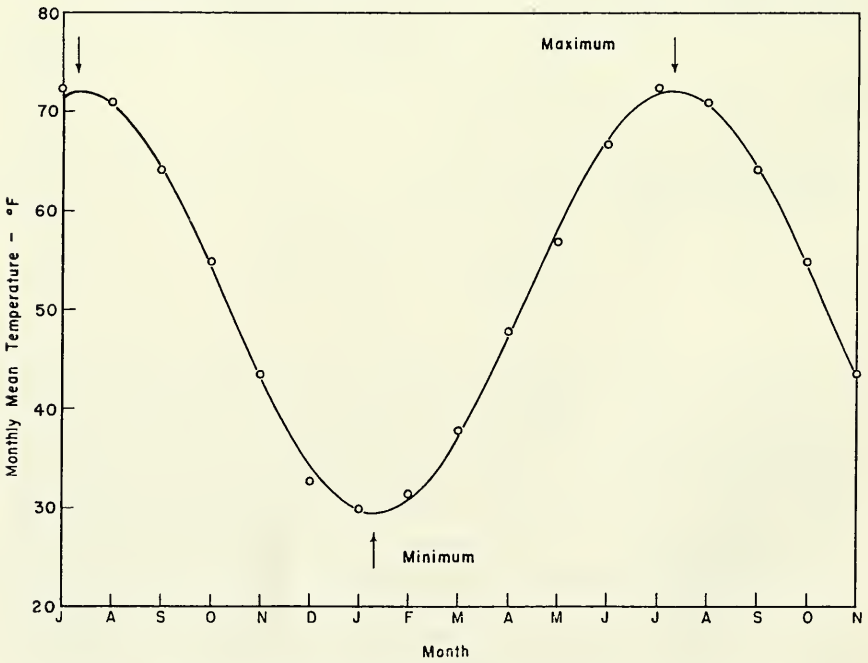

Figure 3. Monthly mean temperatures from Table 2 fitted with a sine curve.

TABLE 2. Monthly mean temperatures in °F. at the Weather Bureau Station in New Haven, Conn., from July 1943 through June 1957.

Year	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	T_r
1943-44	72.6	70.9	63.5	52.6	41.2	29.2	30.8	29.0	34.4	45.0	61.3	67.4	597.9
1944-45	74.2	73.2	65.6	52.7	41.8	29.5	23.3	29.6	44.0	52.2	55.0	67.0	608.1
1945-46	71.6	69.4	67.0	52.6	43.7	27.0	28.8	28.1	43.6	46.2	56.2	64.7	598.9
1946-47	70.4	67.5	65.6	56.8	47.0	34.2	32.7	27.7	35.6	46.2	55.0	63.9	602.6
1947-48	72.5	72.1	64.4	57.5	41.6	30.4	22.8	25.4	36.6	46.2	55.8	65.0	590.3
1948-49	72.0	71.8	64.9	53.8	48.0	34.0	35.7	34.8	38.8	49.6	58.0	69.7	631.1
1949-50	75.1	72.1	62.7	58.1	42.6	35.0	36.0	29.3	34.3	44.7	54.8	65.2	609.9
1950-51	70.7	70.0	61.2	55.9	45.1	33.1	32.2	33.5	39.2	49.7	58.7	65.6	614.9
1951-52	72.4	71.0	64.4	55.1	40.4	35.3	32.9	33.3	37.3	50.5	56.6	68.8	618.0
1952-53	75.0	71.5	64.8	51.7	44.5	35.3	34.2	35.4	39.1	48.6	58.6	67.2	625.9
1953-54	72.1	70.0	65.9	55.0	44.5	37.6	26.4	36.7	38.3	48.7	55.2	66.8	617.2
1954-55	70.7	68.9	62.6	57.5	43.1	33.7	28.6	31.2	37.2	49.1	59.7	65.3	607.6
1955-56	75.4	74.7	63.9	55.8	41.6	26.8	30.2	32.9	33.8	44.1	53.6	67.0	599.8
1956-57	69.4	69.9	61.4	53.1	43.5	36.6	24.3	33.5	38.7	48.4	57.7	69.9	606.4
T_t	1014.1	993.0	897.9	768.2	608.6	457.7	418.9	440.4	530.9	669.2	796.2	933.5	8528.6

The Fourier Series

The plotted means may not define as symmetrical a relation as the sine curve. By Fourier analysis we can add the higher harmonics, corresponding to 2, 3, 4 or more complete cycles in the basic interval covered by one cycle of the sine curve. If we add enough terms the computed curve will fit any observed series exactly, but the equation then has little meaning either biologically or climatologically. Our objective is to add no more terms than are needed to reduce the variance from the scatter of $\bar{y}_t$'s about the fitted line to the same magnitude as the residual error. We may stop well short of this if the scatter seems essentially random even though its variance is significantly larger than the residual variation.

The sine curve in Equation 6 is extended with additional terms to

$$Y = a_o + a_1u_1 + b_1v_1 + a_2u_2 + b_2v_2 + a_3u_3 + b_3v_3 + \ldots \qquad (9)$$

where $u_2 = \cos(2ct)$, $v_2 = \sin(2ct)$, $u_3 = \cos(3ct)$, $v_3 = \sin(3ct)$, etc. and each pair of coefficients a_i and b_i is computed with Equations 7 or 8, replacing u_1 and v_1 by u_i and v_i for $i = 1, 2, 3 \ldots$ successively. The u_i's and v_i's convert the scale of t to orthogonal units in which $\Sigma(u_iv_i) = \Sigma(u_iu_j) = \Sigma(v_iv_j) = 0$ where $i \neq j$. There is the additional advantage that for any given k, $\Sigma u_i^2 = \Sigma v_i^2 = \frac{1}{2}$ k for all values of i, except the last term where k is even and then $\Sigma u_i^2 = k$. The values of u_i and v_i for the first terms of the Fourier series are given in Table 1 for k = 4, 6, 7, 8, 12 and 24 subdivisions per cycle.

A seasonal trend which is not a simple sine curve occurs in the iodine value of butterfat at five stations in central Alberta, Canada, as reported by Wood (1956). Each entry in Appendix Table 1 represents duplicate analyses of the weekly samples of butter in each month for two years beginning in April 1952, or an average of 17.3 determinations. Both the annual total for each station and the month with the peak reading tended to shift in going south from Edmonton to Calgary. According to Wood, the monthly readings in the two years, which have been averaged, did not differ significantly. Although a shift in the phase angle from one location to another accounts for part of the complexity of the average curve, the iodine values for each location could not be fitted adequately with a separate sine curve.

From the sums of products of T_t with the cosines (u_i) and sines (v_i) in Table 1 for the first three harmonics, the seasonal trend of the means in the upper part of Figure 4 is reproduced quite faithfully by the equation:

$$Y = 36.955 + 0.4091u_1 + 1.7318v_1 + 0.0700u_2 - 0.5542v_2 + 0.2233u_3 + 0.7467v_3$$

This curve is merely the overall mean, $a_0 = 36.955$, plus the deviations for each harmonic in each month, as the reader may verify from the last three rows of Appendix Table 1. The Fourier terms have been plotted separately in the lower part of Figure 4 as deviations from the mean a_0, where it is evident that they define successively 1, 2, and 3 complete cycles within the year.

In the present case the biological implications of the successive harmonics are by no means clear. Iodine values are indicative of the unsaturated fatty acid content of butter and are expected to be high during the grass feeding season in May. As noted by the author, the peak in August and September, most pronounced in the North and decreasing southward, was unexpected. Although the biological information gained in fitting a Fourier series is here questionable, the example has served its primary purpose of demonstrating that an apparently irregular curve can be fitted by harmonic analysis with a limited number of constants.

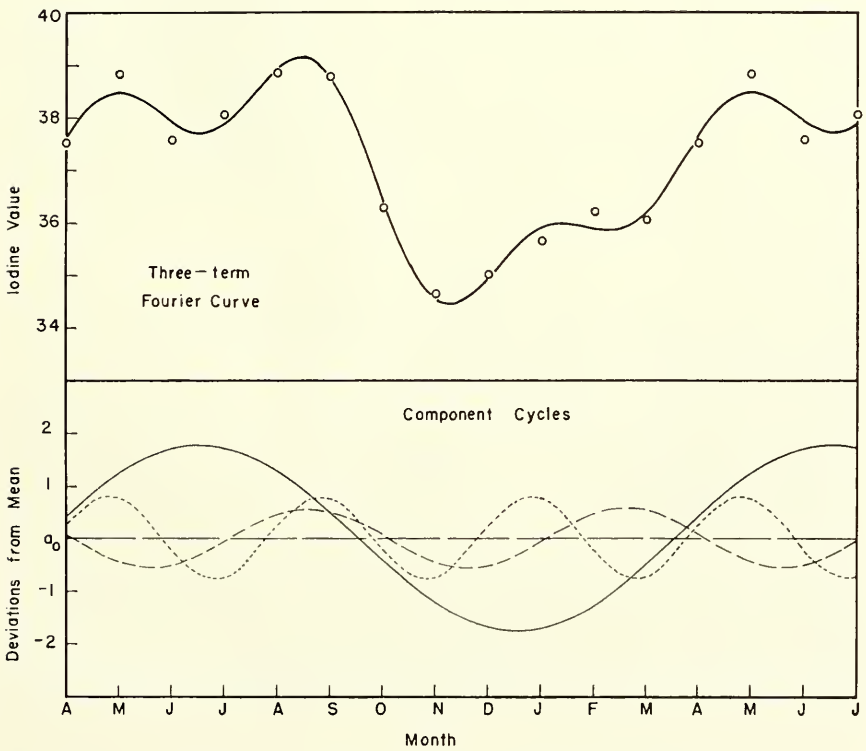

Figure 4. Mean monthly iodine values for butterfat from Appendix Table 1. The sum of the deviations in the lower three curves, added to the mean (a_0), yields the three-term Fourier curve in the upper diagram.

Analysis of Variance

The analysis of variance has the same function in periodic regression as in many other regression problems. The variation in y about the fitted curve is assumed to be normally distributed, equally variable over the length of the cycle, and with deviations independent of each other. The selection of a suitable transform may aid materially in achieving these objectives, as we shall see in a later section. A more troublesome problem is the potential dependence between successive observations through a cycle. Despite the formal analogy of a cross-classification to randomized blocks, the responses in each row represent an ordered sequence rather than an arrangement upon which treatments have been superimposed at random.

One approach is to fit a Fourier series to the column means and compute a serial correlation coefficient from the successive residuals, as described by Anderson and Anderson (1950). In a time sequence, such as of weather records or of attack rates by a contagious disease, these correlations are often significant. An alternative approach, more consonant with the analysis of variance, is to fit a separate Fourier series with a limited number of terms to each replicate. The interaction of rows by columns, or of replicates by periods, is then subdivided into as many parts as may be needed to remove the systematic difference between the trend in each series and the mean trend. In this way we may separate the composite interaction into cyclic trends and residual error. The same argument holds, of course, whether replicates represent successive cycles, such as the years in Table 2, or sampling locations as in Appendix Table 1. The more nearly these separate curves define the periodic trend in each replicate with the fewest terms, the more nearly will the residual error provide an unbiassed estimate of the random error.

With an orthogonal design, the calculation is very similar to that for randomized blocks. The sum of squares between the f totals T_r for replicates, representing successive complete cycles or different locations, corresponds to variation in the statistic a_0 of our separately fitted series. When these totals suggest a trend, we may wish to isolate its linear and quadratic terms to test its form and significance. The sum of squares between the k totals T_t for each interval within the cycle may be subdivided progressively, beginning with a_1 and b_1 for the first harmonic with two degrees of freedom, and following with the second and higher harmonics from the Fourier series, until the scatter about the fitted curve contains no element which we can isolate with profit.

The remaining sum of squares, the interaction of replicates by measured intervals within the cycle, includes not only the random error but also the variation from replicate to replicate of each harmonic in the Fourier series, in so far as these represent systematic rather than random deviations. The

differences between cycles in the first harmonic, with 2(f-1) degrees of freedom, almost certainly should be isolated and tested. In deciding how much farther to partition the interaction, our most useful guide, when available, is the theoretical or expected variance, with which we can compare each mean square. In its absence, we may subdivide the interaction into as many additional terms of the Fourier as have proved useful in fitting the means of all replicates. This rule is rough at best, since a significant higher term may repeat itself so consistently in all replicates that it will not remove a systematic component from the interaction. Alternatively, systematic trends in the individual cycles, corresponding to the second or higher harmonic, may cancel one another when averaged over all replicates.

Mathematical model

These relations may be reduced to more concrete terms by an explicit mathematical model. A single variate occurring in the i^{th} year (or replicate) and the j^{th} month (or interval) is potentially the sum of a number of elements. An element with a subscript i has the same value through a given year but may vary from year to year; an element with a subscript j has a fixed value for a given month but may vary from month to month; an element with both subscripts is specific for a given month and year. With this notation each individual variate y_{ij} may consist of the following terms

$$y_{ij} = (\mu + r_i) + (\alpha_1+a_1'_i)u_{1j} + (\beta_1+b_1'_i)v_{1j} + (\alpha_2+a_2'_i)u_{2j} + (\beta_2+b_2'_i)v_{2j} + t_j + e_{ij} \qquad (10)$$

where the Latin and Greek terms in parentheses correspond to the expectations for the successive statistics of the two-term Fourier curve in Equation 9 for the year i, and $(t_j + e_{ij})$ represents the difference between the observed value y_{ij} and its expectations Y_{ij}. Greek letters stand for the expected values of the same curve fitted to the monthly means over all years or replicates, t_j is the difference between the observed and expected mean for a given month (or other interval), and e_{ij} is the inescapable normal random component.

Our null hypothesis is that each intermediate element in Equation 10 (except the cosines and sines) is zero, which, if true, would then reduce to $y_{ij} = \mu + e_{ij}$. If a single two-term Fourier equation, estimated from the total T_t for each month (or other interval), were to describe the phenomena adequately, our model would simplify to

$$y_{ij} = \mu + \alpha_1 u_{1j} + \beta_1 v_{1j} + \alpha_2 u_{2j} + \beta_2 v_{2j} + e_{ij}$$

all other elements being indistinguishable from a true value of zero. A significant variation of the monthly means about this curve would require

the term t_j, which might represent a third or higher term in the Fourier series or discrepancies common to each replicate year from some other source. All remaining elements, with subscripts i, measure the differences from year to year (or replicate to replicate) in successive terms of the Fourier equation.

Calculation

When the elements in Equation 10 are rearranged in the order in which their variation is isolated in successive rows of the analysis of variance, we have

$$\begin{array}{lccccccccc} y_{ij} = & \mu & + & r_i & + & (a_1u_{1j}+\beta_1v_{1j}) & + & (a_2u_{2j}+\beta_2v_{2j}) & + & t_j \\ \text{Row} & 9 & & 1 & & 2 & & 3 & & 4 \end{array} \quad (11)$$

$$\begin{array}{ccccccc} + & (a_1'_iu_{1j}+b_1'_iv_{1j}) & + & (a_2'_iu_{2j}+b_2'_iv_{2j}) & + & e_{ij} \\ & 5 & & 6 & & 7 \end{array}$$

Separate sums of squares are attributable to the unique combinations of these elements enclosed by parentheses in Equation 11, the number beneath each term identifying the row in the analysis of variance. Their practical calculation is outlined in the workform of Table 3, which may be reduced to that for a sine curve by omitting rows 3 and 6, or extended with additional Fourier terms. Square brackets [ ] designate the sum of the squares or products of the factors they enclose measured from their respective means as the origin, i.e. $[y^2] = \Sigma(y-\bar{y})^2 = \Sigma y^2 - \Sigma^2 y/fk$, or correspondingly $[u_1y] = \Sigma\{(u_1-\bar{u}_1)(y-\bar{y})\} = \Sigma(u_1y)$ since $\Sigma u_1 = 0$. Its other symbols are defined above, in the workform or in Equations 7 and 8. For each sum of products the identity, $\Sigma[u_1y] = [u_1T_t]$, provides a useful check on the arithmetic, which holds similarly for the products with v_1, u_2, v_2, etc. The sum of squares in each row, designated as S_1 to S_{11}, is divided by its degrees of freedom (DF) to obtain the corresponding mean square (MS).

When a given pair of coefficients, a_i and b_i, varies significantly between replicate curves, its harmonic may differ in amplitude, in phase, or in both. Since amplitude and phase angle are computed from non-linear combinations of a_i and b_i, their relative contributions to the sum of squares in row 5 or 6 cannot be separated orthogonally. However, if we disregard phase, we can estimate the total variation in amplitude from replicate to replicate in terms of a single y^2 from $\frac{1}{2}k\Sigma(A-\bar{A})^2$, where A is the semi-amplitude of a given harmonic in a single replicate (Equation 7) and $\bar{A}$ that for the same harmonic in the average curve (Equation 8). For the first harmonic this reduces algebraically to the sum of squares defined in row 10 of Table 3. The difference between this sum of squares and that

TABLE 3. Calculation of the sums of squares (S_i) in the analysis of variance for a two-term Fourier curve.

Row	Term	DF	Sum of Squares, (SS = S_i)
1	Between replicates	$f-1$	$\Sigma T_r^2/k - C_m$
2	Effect of (a_1+b_1)*	2	$\{[u_1T_t]^2 + [v_1T_t]^2\}/\frac{1}{2}fk = S_2$
3	Effect of (a_2+b_2)*	2	$\{[u_2T_t]^2 + [v_2T_t]^2\}/\frac{1}{2}fk = S_3$
4	Scatter about curve	$k-5$	$\Sigma T_t^2/f - C_m - S_2 - S_3$
5	Replicate × (a_1+b_1)*	$2(f-1)$	$\Sigma\{[u_1y]^2 + [v_1y]^2\}/\frac{1}{2}k - S_2 = S_5$
6	Replicate × (a_2+b_2)*	$2(f-1)$	$\Sigma\{[u_2y]^2 + [v_2y]^2\}/\frac{1}{2}k - S_3$
7	Replicate × scatter	$(f-1)(k-5)$	By difference
8	Total	$fk-1$	$\Sigma y^2 - C_m = [y^2]$
9	Correction for mean	1	$\Sigma^2 y/fk = C_m$
10	Replicate × A_1	$f-1$	$2S_2 + S_5 - 2\sqrt{S_2}\,\Sigma\sqrt{[u_1y]^2 + [v_1y]^2}/\sqrt{\frac{1}{2}fk}$
11	Replicate × $Phase_1$	$f-1$	$S_5 - S_{10}$

* The notation $(a_1 + b_1)$ or $(a_1 + b_1)$ is used here and subsequently to designate the sum of the effects of the harmonic coefficients, *not* the sum of the coefficients directly.

in row 5, $S_5 - S_{10} = S_{11}$, we may attribute to differences in phase. A significant variation in the second harmonic in row 6 can be subdivided similarly.

TABLE 4. Variance components for the expectations of the mean squares (MS) in Table 4, where each MS = S_i/DF.

Row	Expected mean square
1	$\sigma_2 + k\sigma_r^2$
2	$\sigma^2 + \frac{1}{2}k(a_s^2+b_s^2)_1 + f\sigma_t^2 + \frac{1}{2}kf(\alpha_1^2+\beta_1^2)$
3	$\sigma^2 + \frac{1}{2}k(a_s^2+b_s^2)_2 + f\sigma_{rt}^2 + \frac{1}{2}kf(\alpha_2^2+\beta_2^2)$
4	$\sigma^2 + f\sigma_{rt}^2$
5	$\sigma^2 + \frac{1}{2}k(a_s^2+b_s^2)_1$
6	$\sigma^2 + \frac{1}{2}k(a_s^2+b_s^2)_2$
7	σ^2

Tests of significance

From our model in Equation 11, the mean square in each row of the analysis of variance contains potentially the variance components in Table 4, on the assumption that each source of variation about the average Fourier curve can be considered a random variable. Replicates, for example, are assumed to be equivalent to a random sample of complete cycles, and the variation of replicates by each term in the Fourier series to represent similarly a random selection. We will further assume that any correlation between successive observations within a replicate is removed in the interaction of replicates by a_1 and b_1 and by a_2 and b_2 in rows 5 and 6 of Table 3, where the effect of each pair of coefficients is symbolized as "(a_1+b_1)", "(a_2+b_2)", etc.

Under these assumptions, the variance components are essentially the same as those for other replicated regressions, whether linear, curvilinear or harmonic. The components for regression from a_i and b_i in Equation 11 are designated as a_s^2 and b_s^2 in Table 4 and converted to units of y^2 by the factor $\frac{1}{2}k \doteq \Sigma u_i^2 = \Sigma v_i^2$. The variance components σ^2 with subscripts for replicates (r) and time (t) are already in units of y^2, as is the random variance σ^2 which recurs in each MS and may be an undivided composite.

The error variance for a test of significance or a measure of precision depends upon which of the relevant components in Table 4 differ effectively from zero. It may be a single mean square or a linear combination of variances, and will frequently be designated as s^2. When testing the null hypothesis that the additional component is zero in the mean square V_i in

row i = 1, 4, 5 or 6, the appropriate s^2 is V_7. The significance of each observed $F = V_i/V_7$ is determined by reference to a table of F or the variance ratio, such as that given by Fisher and Yates (1957) or by Pearson and Hartley (1954). If the mean square for scatter about the fitted average curve in row 4, for example, is significantly larger than that for the residual variation in row 7, we would conclude that the deviations about the replicate curves have a common element.

An F test of the Greek coefficients in rows 2 and 3 is more involved. If the scatter in row 4 or the interaction in row 5 or 6 should prove less than or negligibly larger than the random error, its component would drop out of the sum in row 2 or 3 of Table 4, and the remaining components would determine which single mean square is the appropriate error. When both the scatter in row 4 and the interaction of the first or second harmonic with replicates are significant, the appropriate error is a linear combination of the mean squares (V_i) in three different rows (Anderson and Bancroft, 1952). For the effect of $(a_1 + b_1)$, the error is $s^2 = V_4 + V_5 - V_7$ with approximately n′ degrees of freedom, estimated as

$$n' = \frac{(V_4 + V_5 - V_7)^2}{(V_4^2/n_4) + (V_5^2/n_5) + (V_7^2/n_7)} \qquad (12)$$

For an approximate test of significance, we refer

$$F' = V_2/(V_4 + V_5 - V_7) \qquad (13)$$

to a table of the variance ratio (F) with $n_1 = 2$ and $n_2 = n'$ degrees of freedom. Similarly, for the second term in the Fourier series the error is $s^2 = V_4 + V_6 - V_7$, with F′ and n′ determined by Equations 12 and 13, replacing subscript 5 by subscript 6.

Examples

The analysis of variance in Table 5 has been computed from the monthly mean temperatures at New Haven in Table 2. An inspection of the yearly or replicate totals reveals no obvious trend, except possibly for a series of warmer years in the middle of this 14-year period. Since a parabola fitted to the T_r's (not shown here) did not approach significance, we will consider the differences in T_r a random variable. Their mean square V_1 exceeds the interaction V_7 significantly ($P < 0.02$). The sine curve for the monthly totals (T_t) accounts for 96.9% of the total sum of squares and is obviously highly significant. Although the mean square for the second term in the Fourier series, $(a_2 + b_2)$, is larger than the scatter around the two-term Fourier curve, its error depends upon the significance of the mean squares in rows 4 and 6.

TABLE 5. Analysis of variance of the monthly mean temperatures at New Haven, Conn., in Table 2.

Row	Term	DF	SS	MS	F
1	Between years	13	138.95	10.689	2.19
2	Months, effect of (a_1+b_1)	2	38026.32*	19013.160	1695
3	" effect of (a_2+b_2)	2	40.07	20.035	2.20
4	" scatter	7	57.76*	8.252	1.69
5	Years × Month (a_1+b_1)	26	291.71	11.220	2.29
6	" × " (a_2+b_2)	26	237.17	9.122	1.87
7	" × " scatter	91	445.08	4.891	
8	Total	167	39237.06		
9	Correction, C_m	1	432958.44		
10	Year × $Amplitude_1$	13	165.55	12.735	2.60
11	Year × $Phase_1$	13	126.16	9.705	1.98
12	Year × $Amplitude_2$	13	132.05	10.158	2.08
13	Year × $Phase_2$	13	105.12	8.086	1.65

* When recomputed with $\Sigma u_1^2 = \Sigma v_1^2 = 5.999824$ instead of their expectations, $\frac{1}{2}k = 6$, these SS were corrected to 38027.43 and 56.64 respectively, no others differing by more than 0.01.

When compared with the interaction V_7, both the first and second Fourier terms varied significantly from year to year, but the scatter about the average curve in row 4 fell within the acceptable range. This last result is in line with Craddock's finding (1955) that temperature records in the northern hemisphere agree quite generally with a two-term Fourier series. Both the scatter in row 4 and its interaction with years in row 7 might have been subdivided by adding a third term to the Fourier series, as in fact was done, but without a significant reduction in the remaining mean squares. Since V_4 is not significant, we may retain our null hypothesis that its variance component σ_t^2 is zero, and compare the mean squares from the first and second terms in the Fourier curve for the 14-year average with their respective interactions by years. For $(a_2 + b_2)$, we have $F = 20.305/9.122 = 2.20$, which is not significant.

To separate the differences in amplitude and in phase, the variations of the Fourier curve from year to year in rows 5 and 6 have been subdivided in the last four rows of Table 5. These indicate that for both the first and second harmonic, the amplitude or annual range differed somewhat more from year to year than the phase or date of the maximum. The variation in the mean monthly temperature will be considered later in more detail.

TABLE 6. Analysis of variance of the average monthly iodine values in Appendix Table 1.

Row	Term	DF	SS	MS	F	F′
1	Place	4	65.8543	16.4636	141.20	
2	Months, (a_1+b_1)	2	94.9890	47.4945		18.56†
3	" (a_2+b_2)	2	9.3625	4.6812		6.65††
3′	" (a_3+b_3)	2	18.2217	9.1108	17.74	
4	" scatter	5	2.5673	.5135	4.40	
5	Place × Month (a_1+b_1)	8	17.2916	2.1614	18.54	
6	" × " (a_2+b_2)	8	2.4569	.3071	2.63	
7	" × " scatter	28	3.2652	.1166		
10	Place × $Amplitude_1$	4	5.0955	1.2739	10.93	
11	" × $Phase_1$	4	12.1961	3.0490	26.15	

† $s^2 = 2.5583$, $n' = 10.27$; †† $s^2 = 0.7040$, $n' = 7.62$.

From the analysis in Table 6 of the iodine values in Appendix Table 1, the three-term Fourier curve accounts for 97.9% of the variation between the monthly totals; there would be little point in adding more terms to the series. The five creameries or replicates differed very significantly in their means and in the first harmonic $(a_1 + b_1)$. When the latter (row 5) was subdivided between amplitude and phase (rows 10 and 11), differences in phase proved the more important. The interaction of place with the third and higher terms proved so nearly equal that they have been pooled in estimating the random error in row 7. From its variance components, the error for testing (a_3+b_3) in row 3′ is the mean square in row 4. Since all random components in the mean squares for (a_1+b_1) and (a_2+b_2) are significant, each is tested in terms of F′. For the first term, $F' = 47.4945/(0.5135 + 2.1614 - 0.1166) = 18.56$ and the divisor (2.5583) has approximately $n' = 2.5583^2/(0.5135^2/5 + 2.1614^2/8 + 0.1166^2/28) = 10.27$ degrees of freedom by Equation 10, and for the second term $F' = 6.65$ with $n' = 7.62$. All three terms of the curve plotted in Figure 4 are clearly significant.

A systematic trend from replicate to replicate may be illustrated by the progressive change in the standing electrical potential (Burr, 1945) of an elm tree, which varies diurnally. The hourly potentials, as read from the daily record for eight three-day periods from August 1 to 25, 1953, have been coded in Appendix Table 2 for ease of analysis. The hourly means (in code) have been fitted with the two-term Fourier curve (Equation 9):

$$Y = 49.964 - 6.605u_1 - 15.084v_1 + 1.357u_2 + 1.146v_2$$

Decoded, the estimated average potential for each hour is

$$Y' = -66.654 + 2.202u_1 + 5.028v_1 - 0.452u_2 - 0.382v_2$$

which has been plotted as the solid curve of Figure 5. Except for a slight flattening at the upper and lower limits, as if limited by maximal and minimal potentials, the fit seems very good; how good we can determine from the analysis of variance in Table 7.

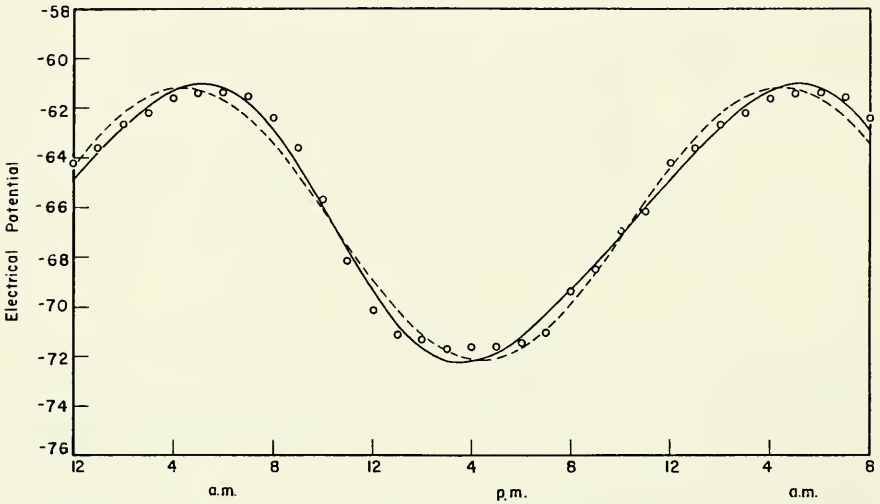

Figure 5. Mean hourly potentials in an elm tree fitted with a two-term Fourier curve (solid line) and with a sine curve (broken line), from Appendix Table 2.

Over this period of 25 days, the average potentials, all initially negative, decreased progressively, as indicated by the rise in T_r in Appendix Table 2. In consequence, the variation between replicates has been subdivided into a highly significant linear trend and the scatter about this trend, in rows 1 and 1′, with the latter still much greater than the random error in row 7. This trend was succeeded toward the end of the month by a drastic change in the diurnal pattern, possibly in response to the prolonged dry spell in that August.

Since the mean squares for both the first and second Fourier terms are so much larger than the remaining variation between the hourly means (row 4), the two-term Fourier curve seems to fit the plotted points in Figure 5 better than the simpler dotted sine curve. However, the interaction of replicates by the first and by the second term both exceed the residual variation so considerably that the significance of the mean squares in rows 2 and 3 must be tested by F′. By this criterion, the first term or sine curve

TABLE 7. Analysis of variance of the tree potentials for the eight 3-day periods in Appendix Table 2.

Row	Variance due to	DF	SS	MS	F, F′
1	Linear trend on periods	1	20259.80	20259.796	21.37
1′	Scatter about trend	6	5687.82	947.971	176.33
2	Hours, (a_1+b_1)	2	26030.12	13015.062	66.17†
3	" (a_2+b_2)	2	302.80	151.398	1.78††
4	" scatter	19	297.20	15.642	2.91
5	Period × Hour (a_1+b_1)	14	2609.93	186.424	34.68
6	" × " (a_2+b_2)	14	1048.06	74.861	13.92
7	" × " scatter	133	715.01	5.376	
8	Total	191	56950.74		
10	Period × Amplitude_1	7	1822.73	260.390	48.44
11	" × Phase_1	7	787.20	112.457	20.92

† $s^2 = 196.690$, $n' = 15.50$; †† $s^2 = 85.127$, $n' = 17.53$.

is highly significant but not the second term ($F' = 1.78$, $P \doteq 0.20$). Despite its apparently better fit, the more complex curve offers no real advantage in describing the average diurnal variation in tree potential. As judged from Table 7, in studying the relation between the daily tree potentials and environmental factors, such as temperature, cloudiness, soil moisture and humidity, the hourly readings for each day might well be replaced by the first five constants in a Fourier series (a_0, a_1, b_1, a_2 and b_2) and these used as the dependent variables in a comprehensive analysis.

Transformations of the Variate

In meeting the assumptions of the analysis of variance, the adoption of a suitable unit for the response is often critical. An unsuitable original measurement or count can often be transformed to a unit which is either additive or has a variance independent of the mean. In fulfilling one requirement we frequently meet or approximate the other assumptions in the analysis of variance, and in some cases acquire an expected variance, with which the observed variation can be compared.

Sometimes the transformation can be based upon past experience with the variate or upon a biological relation. Thus, if we expect our measurement to change proportionately or percentagewise with time, such as the incidence of a contagious disease, the appropriate unit would be the loga-

rithm of the incidence. If the initial variable is the number of occurrences or individuals in each unit of time, its distribution, apart from the periodic effect, may well be Poisson. The expected variance of each Poisson count is its unknown population mean, but the appropriate transform, the square root of each count, has a constant variance of 0.25. Our data may be binomial percentages which can be assumed to measure indirectly an underlying threshold response, some function of which is normally distributed in the biological population. The additive transform is then the probit, or the unit, usually the logarithm, to which the probit is linearly related.

Log-transforms

Since the logarithms of many biological measurements are normally distributed, the logarithmic transformation should be of equal value in periodic regressions, such as of contagious diseases in animals and plants. An example from man is the seasonal variation in the death rate from pneumonia, as recorded in the monthly reports of the Metropolitan Life Insurance Company (1945-1955). The month of September, when deaths are near a minimum, has been selected here as the starting time (t_0) for each annual cycle in Appendix Table 3, where each monthly rate per 100,000 has been transformed to its logarithm, a unit which stabilizes the variance through the year. The log-death rates for September 1945 through December 1949, when deaths were classified by the 5th Revision of the International List of Causes of Death, have been adjusted here to conform with the 6th Revision used subsequently by subtracting from each earlier

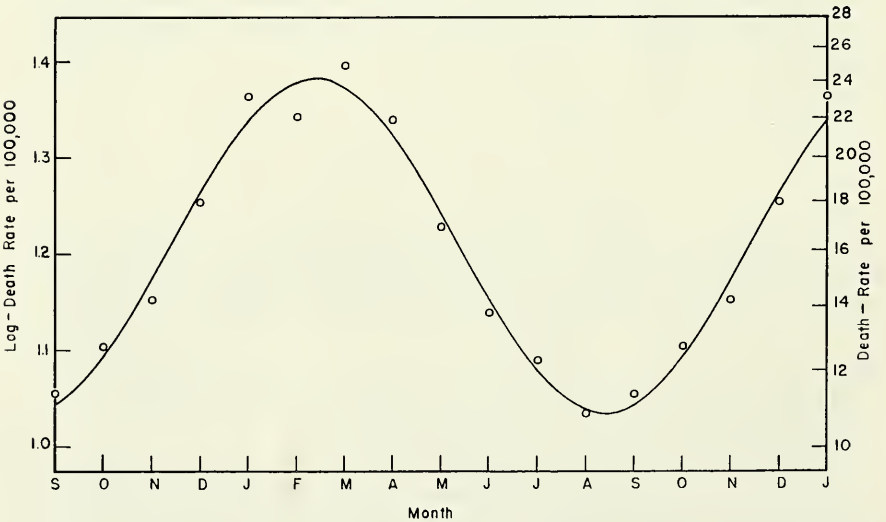

Figure 6. Mean monthly log-death rates from pneumonia and fitted sine curve, from Appendix Table 3.

log-death rate the mean difference (0.235) during the twelve months of 1950 when both criteria were reported.

The sine curve, $Y = 1.2087 - 0.1647u_1 + 0.0535v_1$, has been computed with Equation 8 from the monthly totals T_t and plotted in Figure 6. By Equation 4 the seasonal range in the mean log-death rate is more than two-fold, $2A = 0.3464 = \log(2.220)$. By Equation 5 and Figure 2, its maximum at $\tan \theta' = 0.32507$ and phase angle $\theta = 2.8273$ radians, corresponds to 5.400 months from the starting point of each annual cycle in mid-September. This places the maximum death rate at approximately February 25 and the minimum six months later.

TABLE 8. Analysis of variance of the log-death rates from pneumonia in Appendix Table 3.

Row	Term	DF	SS	MS	F, F′
1	Years, trend on x_1	1	.80964	.80964	246.09
1′	" trend on x_2	1	.06494	.06494	19.74
1″	" scatter	7	.02303	.00329	1.52
2	Months, (a_1+b_1)	2	1.79995	.89998	107.83†
3	" (a_2+b_2)	2	.00445	.00223	0.25††
4	" scatter	7	.03650	.00521	2.40
5	Years × Month (a_1+b_1)	18	.09540	.00530	2.44
6	" × " (a_2+b_2)	18	.10607	.00589	2.72
7	" × " scatter	63	.13662	.00217	
10	Year × A_1	9	.06990	.00777	3.58
11	" × $Phase_1$	9	.02550	.00283	1.31
12	" × A_2	9	.07347	.00816	3.76
13	" × $Phase_2$	9	.03260	.00362	1.67

† $s^2 = 0.008346$, $n' = 12.6$; †† $s^2 = 0.008939$, $n' = 13.6$.

The progressive decrease in the yearly totals (T_r) (Appendix Table 3) has been fitted with the linear and quadratic orthogonal polynominals, x_1 and x_2, for a series of 10 (Fisher and Yates, 1957). This parabola accounts effectively (97.4%) for the trend between years, as judged from rows 1 to 1″ of the analysis of variance (Table 8). A similar proportion (97.8%) of the sum of squares between the monthly totals (T_t) is absorbed by the harmonic coefficients a_1 and b_1. Since the mean square for the second harmonic is less than that for the remaining scatter, little would be gained by adding more terms.

The variation from year to year in both of the first two harmonics exceeds the remaining interaction with years significantly, despite the disappearance of the 2nd harmonic from the average curve. When isolated from the residual sum of squares in row 7, the mean squares for the higher terms decreased progressively, but in the absence of an expected error variance with which to compare them, they have been pooled in the analysis. As judged from the last four rows in Table 8, the first two harmonics were considerably more stable in phase from year to year than in amplitude.

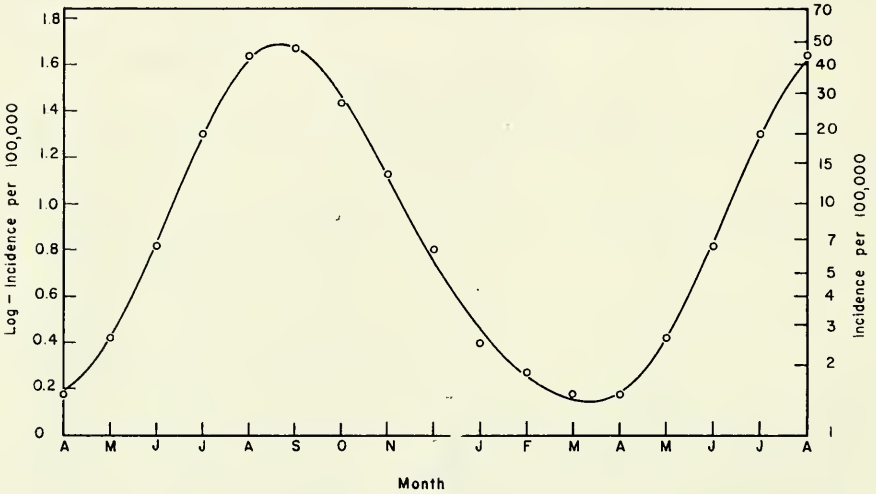

Figure 7. Mean monthly log-incidence of poliomyelitis in the United States with two-term Fourier curve, from Appendix Table 4.

A similar analysis of another contagious disease with a marked seasonal incidence, poliomyelitis, reveals a different pattern. The U. S. monthly incidences per million have been changed to logarithms in Appendix Table 4 (Serfling and Sherman, 1953, 1958) and analyzed in Table 9. Although a parabola accounts for much of the overall difference between years (T_r), the scatter about this trend (row 1″) is here far larger than that about the annual curves (row 7). Instead of a simple sine curve, the monthly totals (T_t) define the two-term Fourier curve in Figure 7, with both terms significant and the equation

$$Y = 1.8517 - 0.6397u_1 + 0.4161v_1 - 0.0252u_2 - 0.0861v_2$$

This increases in 24 weeks from a minimum, approximately on March 23, to a peak 35 times as great on September 7, and then returns in the following 28 weeks to its minimum. Here the variation in both terms from year to year is about equally distributed between amplitude and phase.

TABLE 9. Analysis of variance of seasonal incidence of poliomyelitis in Appendix Table 4.

Row	Term	DF	SS	MS	F, F′
1	Years, linear trend	1	3.49183	3.49183	10.34
1′	" quadratic curv.	1	2.86902	2.86902	8.49
1″	" scatter	12	4.05418	.33785	94.45
2	Months, (a_1+b_1)	2	52.40486	26.20243	377.83†
3	" (a_2+b_2)	2	.72484	.36242	11.55††
4	" scatter	7	.14494	.02071	5.79
5	Years × Month (a_1+b_1)	28	1.46223	.05222	14.60
6	" × " (a_2+b_2)	28	.39758	.01420	3.97
7	" × " scatter	98	.35051	.00358	
8	Total	179	65.89999		
10	Years × $Amplitude_1$	14	.71986	.05142	14.37
11	" × $Phase_1$	14	.74237	.05303	14.82
12	" × $Amplitude_2$	14	.19934	.01424	3.98
13	" × $Phase_2$	14	.19824	.01416	3.96

† $s^2 = 0.06935$, $n' = 30.29$; †† $s^2 = 0.03137$, $n' = 14.35$.

Square root transform

The advantages of a theoretical error term are evident in the square root transformation for a Poisson variate. Data on the number of normal human births per hour have been assembled by King (1956) from the records of five hospitals, the two with the fewest births having been combined in Appendix Table 5 into a single series (A). If the number of births per hour within each series had varied entirely at random, we would expect its 24 values to follow the Poisson distribution and its variance to equal its mean. Because of differences in the size of the four series and potentially in the hour of birth, the variance has been stabilized by transforming each number of births, ranging from 153 to 508, to its square root (Bartlett, 1936). The hourly means have been plotted in Figure 8 and fitted with the sine curve, $Y = 18.3542 + 0.1085u_1 + 1.3615v_1$.

The adequacy of a simple sine curve has been tested by the analysis of variance in Table 10 of the transformed variates y. If our Poisson hypothesis is correct, the mean square for error in row 7, $s^2 = 0.242$ with 63 degrees of freedom, should not differ significantly from its expectation 0.25. Since the agreement is excellent, each sum of squares for which s^2

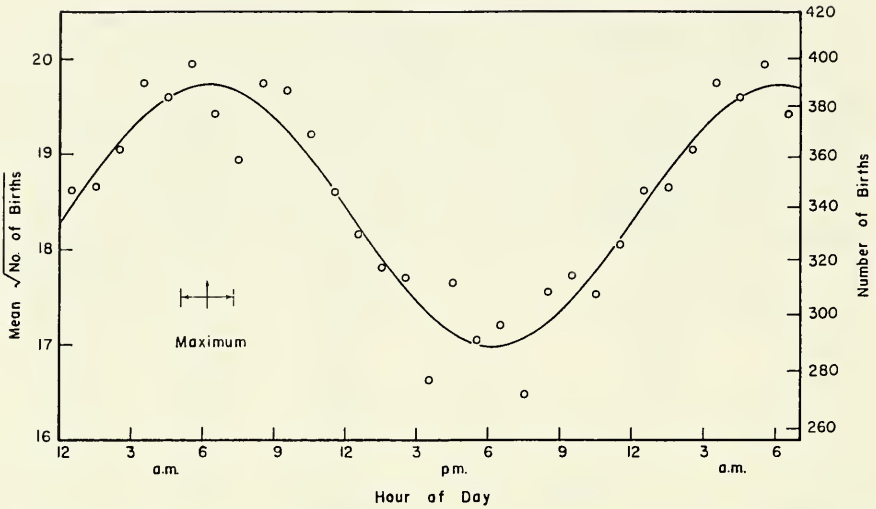

Figure 8. Mean hourly incidence of births in five hospitals and sine curve, from Appendix Table 5.

serves as the error becomes a χ^2 when divided by 0.25 and has the same number of degrees of freedom as before.

The average sine curve in Figure 8 accounts for 89.9% of the variation in the means, with the highest birth rate at 6:12 a.m. Although the remaining scatter is significant ($\chi^2 = 40.42$, P = 0.007), it would not be reduced appreciably by adding the second term in a Fourier series. Separate sine curves for the four series also differed significantly, primarily in

TABLE 10. Analysis of variance of the hourly frequency of human births in Appendix Table 5; χ^2 = SS/0.25.

Row	Term	DF	SS	MS	χ^2	P
1	Between series	3	750.1977	250.0659	3000.79	< .001
2	Hours, effect of (a_1+b_1)	2	89.5420	44.7710†		< .001
4	" scatter	21	10.1042	.4812	40.42	.007
5	Series × Hour (a_1+b_1)	6	7.2891	1.2148	29.16	< .001
7	" × " scatter	63	15.2561	.2422	61.02	.55
8	Total	95	872.3891			
10	Series × Amplitude$_1$	3	5.3943	1.7981	21.48	< .001
11	" × Phase$_1$	3	1.8948	.6316	7.58	.055

† $F' = 44.7710/1.4538 = 30.80$, $n_1 = 2$, $n' = 8.2$.

amplitude and relatively little in phase. The larger deviations in birth time, or its recording, in row 4 tend to recur in all four series, due in part, King suggests, to similarities in hospital routine. Thus the recording of births may be delayed by the nurses' conference between 7 and 8 a.m. when the staff changes, and the balanced low and high points in the hours starting at 3 and 4 and at 7 and 8 p.m. may have similar explanations. This location of observation periods when the recording may be at fault is another advantage of periodic regression. Because of the significant variance components in rows 4 and 5, the critical test for $(a_1 + b_1)$ in the average sine curve is $F' = 30.80$ with an error variance of $s^2 = 1.4538$ $(n' = 8.19)$ and $P < 0.001$.

Probit transform

In biossays of toxicants, such as insecticides or fungicides, and of drugs, the susceptibility of the test organism varies so commonly and usually so unpredictably that a reference or Standard preparation is almost invariably tested concurrently with the sample or Unknown. The variation in susceptibility may be so large, however, as to complicate the selection of a suitable range of dosage levels, especially when the response is a binomial percentage. In an extreme example, the same series of fungicidal concentrations might kill all test spores at one season and none at another. In either case the experiment would be valueless as an assay. If the spore susceptibility were to vary predictably through the year, the concentrations could be so adjusted as to obtain on each occasion an adequate number of intermediate mortalities between 0 and 100 percent. A response in which the seasonal variation has been studied systematically is that of the toad *Bufo arenarum* to chorionic gonadotrophin (Penhos et al, 1954). For two years 40 male toads were collected in the field on the first of each month and on the following day injected in four lots each of 10 toads with the same four dosage levels of the International Standard. The number of individuals in each lot which reacted positively, by releasing sperm, is recorded in Table 11. Not more than one dose in each test produced a reaction of either 0 or 100 percent.

Our problem is to predict from these data the response to be expected at each dosage level in each month of the year. As an all-or-none reaction, we would expect the probit for each percentage to be linearly related to the logarithm of the dose, as indeed proved true. The first step, therefore, was to convert each percentage between zero and 100 to its empirical probit, and to estimate the provisional slope $b = 5.27$ from these values on the assumption that all 24 curves are parallel. From these parallel preliminary curves a provisional expected probit could be estimated for each lot in which none or all of the toads reacted, and then by suitable tables

TABLE 11. Number of toads, *Bufo arenarum*, in each group of 10 reacting positively to four different doses of chorionic gonadotrophin measured in international units per animal, and the log-ED50 computed from each test and from the average sine curve. (Penhos et al, 1954)

Month	1951-52					1952-53					Log-ED50 from sine curve
	No.(+) at dose				Log-ED50	No.(+) at dose				Log-ED50	
	40	30	22.5	15		40	30	22.5	15		
Nov	10	8	7	3	1.272	10	9	6	2	1.288	1.318
Dec	9	7	5	2	1.358	9	7	6	3	1.325	1.353
Jan	8	6	4	2	1.404	7	5	3	1	1.471	1.402
Feb	7	5	4	1	1.452	8	6	4	1	1.420	1.453
Mar	7	5	3	1	1.464	6	3	2	0	1.552	1.492
Apr	7	4	3	0	1.502	8	6	2	0	1.471	1.508
May	9	2	1	0	1.536	8	4	1	0	1.520	1.498
Jun	9	5	4	1	1.420	3	5	4	1	1.437	1.464
Jul	9	6	4	1	1.405	9	6	3	0	1.439	1.415
Aug	10	7	4	2	1.353	9	7	3	2	1.388	1.364
Sep	10	8	5	3	1.304	10	7	4	2	1.356	1.325
Oct	10	8	5	2	1.322	10	8	4	2	1.339	1.308

(Fisher and Yates, 1957) its corresponding working probit. This completes the set of 24 probits at each of the four dosage levels, their sums leading to a new unweighted provisional slope of b = 5.704. From the sums of the 8 probits for each of the 12 calendar months, a sine curve could be computed by Equation 8 for predicting the mean probit in each month as $Y = 4.9751 + 0.5327u_1 - 0.2693v_1$. With the provisional b and Y it was a simple matter to calculate the expected probit for each of the four dosage levels in each calendar month. These determine the weighting coefficients w and, with the observed proportion of positive reactions in each lot, the working probits y for computing the maximum likelihood estimates of the 24 curves. (Bliss, 1952; Finney, 1952)

The variation in y about the 24 separately computed curves was well within the sampling error, $\Sigma\chi^2 = 15.84$ for 38 degrees of freedom. When tested for differences in slope, the curves proved satisfactorily parallel ($\chi_b^2 = 4.28$, n = 23) with a combined slope of $b_c = 5.4724$. Given this slope and for each curve its weighted mean log-dose $\bar{x}$ and probit $\bar{y}$, the ED50 in logarithms has been determined for each month as listed in the Table 11. The sums of the replicate responses in the two years were then fitted with the single sine curve

$$\text{Log-ED50} = 1.4083 - 0.08987u_1 + 0.04462v_1$$

with which the expectations in the last column of Table 11 have been determined. The computed curve and the observed log-ED50 for each month have been plotted in Figure 9.

In an analysis of variance, the log-ED50's for the two years, agreed in their annual means, in their separately fitted sine curves, and in the random scatter about these curves. An expected variance was then determined for each log-ED50 from the sum of the weights (Σw) for its log-dose probit curve and the square of the difference, $(\bar{y}-5)^2$. These varied by less than 7 percent so that an average variance, $\hat{\sigma}^2 = 0.001795$, could be based upon two means, $\overline{\Sigma w} = 18.775$ and $\overline{(5-\bar{y})^2} = 0.06742$, from the internal evidence of the separate monthly determinations. With this expected error variance, the total sum of squares about the average sine curve (from the analysis of variance of the log-ED50's) could be converted to $\chi^2 = 0.023881/0.001795 = 13.31$ with 21 degrees of freedom.

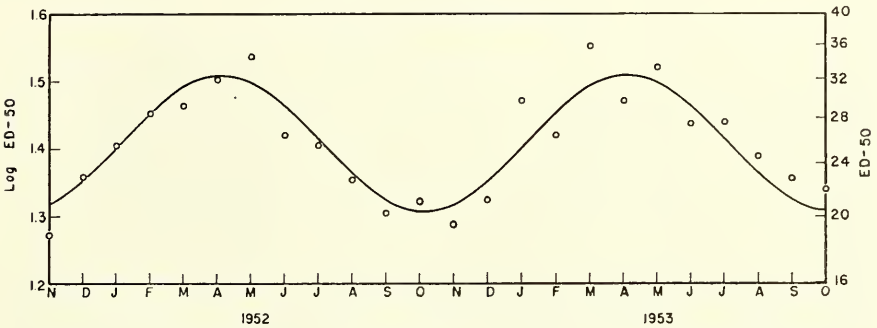

Figure 9. Log-ED50 for gonadotrophin in toads in 24 successive months and annual sine curve, from Table 11.

The observations in Table 11 agree so well with our mathematical model that the three constants in the sine curve plus the combined slope b provide an adequate description of the response of this species to gonadotrophin through the two years of the experiment. Indeed, the three main sources of variation — of the working probits y about the 24 straight lines, between the slopes of these lines, and of the log-ED50's about the sine curve — all had smaller χ^2's than would be expected binomially and were consistent with one another. When totalled over all sources, $\Sigma\chi^2 = 33.425$ with approximately 82 degrees of freedom, after allowing for each probit with an expectation of less than 0.5 positive or negative response. The probability for so small a combined χ^2, $P < 0.000{,}001$, is well outside the range attributable to our initial hypothesis of simple binomial variation.

The seeming paradox can probably be traced to differences in the inherent sensitivity of the field-collected experimental animals. If on a given day these represented several collecting points with unequal thresholds of

response, and if the toads from each location were assigned equally or at random to four test groups, each group of n toads would be a mixture from several populations of sensitivity. For a given dose of hormone, the mean of the p's for the different populations would have an unbiassed proportionate response $\bar{p}$ but, as noted by Kendall (1945), its variance would be reduced from the binomial npq, as assumed in probit analysis, to $n\bar{p}\bar{q} - nV(p)$, where $V(p)$ is the variance in p between populations. Why this would reduce the observed variance may be illustrated by a hypothetical extreme case in which half of the toads at a given dose were collected from a field population of resistant individuals which never reacted and the other half from a different source of very susceptible toads which always reacted. Their combined response would always be exactly 50 percent with a variance of zero.

Adjustment by Covariance

A biological response may be influenced by prior or concomitant variables which, though measurable, are impossible or impracticable to control. A climatic factor, for example, is far easier to measure than to control, and any effect it may have upon a biological response can then be adjusted by covariance. If the covariate is quite unrelated to the cyclic pattern of the response or variate, covariance may reduce the experimental error in the response and strengthen its underlying periodic regression. Alternatively, the covariate may display periodicities so similar to that of the variate, that covariance greatly reduces or eliminates the initial periodicity in the response; it then aids in interpreting the underlying phenomenon. In either case, the adjustment for the covariate depends primarily upon the linear regression of the response y upon the covariate x as computed from the sums of squares $[x^2]$ and of products $[xy]$ in the error row of the analysis.

A case in point is the diurnal variation in the heat exchange of cows reported by Thompson (1954). In an experimental barn under close environmental control, the average heat exchange was determined in BTU's per hour for six animals on each of three days. These measurements were paralleled by a record of the humidity expressed as pounds of water per pound of dry air, the mixing ratio, on the three days of the test, in all cases at an average temperature of 50°F. In fitting a sine curve to the initial data, Thompson noted that the humidities seemed not to follow any periodic pattern. In Appendix Table 6, the individual observations of humidity have been coded and the BTU's transformed to logarithms (−3) with a gain in consistency.

Three columns of the analysis of covariance in Table 12 are sums of squares from analyses of variance of the covariate $[x^2]$ and of the variate $[y^2]$, and the corresponding sums of their products $[xy]$ in which the num-

TABLE 12. Analysis of covariance of data on hourly heat exchange of cows in Appendix Table 6. $B_e^2 = [xy]^2/[x^2]$ in error row 7, $B_s^2 = [xy]^2/[x^2]$ from sums in last rows; the reduced error MS = $\{[y^2] - B_e^2\}$/DF; in rows 1 to 5 each reduced MS = $\{[y^2] - (B_s^2 - B_e^2)\}$/DF.

Row	Term	DF	$[x^2]$	$[xy]$	$[y^2]$	$B_s^2 - B_e^2$	DF	Red.MS	F
1	Between days	2	1.4602	.11230	.009879	.007897	2	.000991	2.33
2	Hours, $(a_1 + b_1)$	2	.0475	.02411	.071887	.005048	2	.033420	5.95††
4	" scatter	21	1.4652	.08474	.038946	.003087	21	.001708	4.02
5	Days × Hour $(a_1 + b_1)$	4	.1989	.04338	.024827	.007500	4	.004332	10.19
7	Error	42	1.1676	.12923	.031714	†.014303	41	.000425	33.65
1+7	Error + days	44	2.6278	.24153					
2+7	" + $(a_1 + b_1)$	44	1.2151	.15334					
4+7	" + scatter	63	2.6328	.21397					
5+7	" + days × $(a_1 + b_1)$	46	1.3665	.17261					

† B_e^2, †† $s^2 = 0.005615$, $n' = 5.14$

$b_{yx} = 0.11068$ from row 5.

bers formerly squared are here cross-multiplied, all other operations being identical. Comparisons of the mean squares from rows 2 and 4 show in the column for $[y^2]$ a well-marked sine curve ($F = 17.17$) in terms of the heat exchange but in that for $[x^2]$ no trace of a sine curve ($F = 0.34$) in terms of the mixing ratio. In consequence, the covariate x is here essentially an environmental rather than an explanatory adjustment. The second term of a Fourier series fitted to the heat exchange proved negligible and has not been isolated in Table 12. In the error row, representing the interaction of days by scatter, the highly significant linear regression of y upon x ($F = 33.65$), accounts for $100 \times 0.014303/0.031714 = 45\%$ of the unadjusted error in the log-BTU, y.

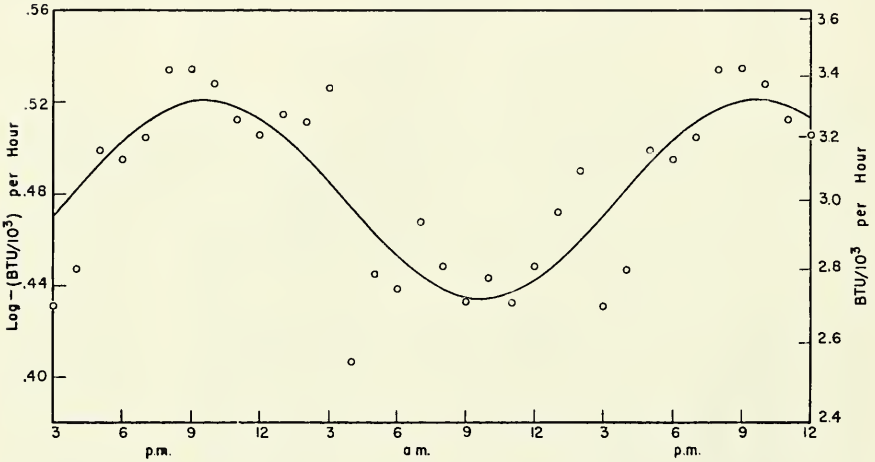

Figure 10. Log-BTU exchange in cows and sine curve for intervals starting at each stated hour, adjusted for differences in relative humidity, from Appendix Table 6.

After correction for the covariate, the ratio of the reduced mean square for the average sine curve in row 2 has increased relative to that for scatter in row 4 ($F = 19.57$). However, both the scatter in row 4 and the interaction of days by $(a_1 + b_1)$ in row 5 are so very significant ($P < 0.001$), that the appropriate error for the average sine curve is the combination of the reduced mean squares in rows 4, 5 and 7, $s^2 = 0.001708 + 0.004332 - 0.000425 = 0.005615$ with 5.14 degrees of freedom, from which $F' = 5.95$ and the true significance of the adjusted curve is $P < 0.05$. The hourly means, adjusted for the covariate x with the slope $b_{yx} = 0.11068$, have been plotted in Figure 10 with the adjusted sine curve, $Y = 0.47793 - 0.00736u_1 + 0.04255v_1$.

Precision of the Computed Curve

The statistics of the Fourier curve, such as its mean amplitude and phase, are estimates subject to error. In considering their precision, we will restrict ourselves to the first harmonic or sine curve. Of the several sources of variation to which it is subject, the most nearly random is the residual error about the series of curves fitted separately to each replicate and designated as σ^2 in Table 4. A second source is the scatter of, say, the monthly means of the f replicates about the average fitted curve, which involves the additional variance component σ_t^2. A third source, the variation between the sine curves fitted to each replicate, is divided between the sum of squares for replicate means or totals (a_o) with f-1 degrees of freedom, and that for the interaction of replicates by (a_1+b_1) with 2(f-1) degrees of freedom. The replicate means especially may include a systematic element which, when segregated, leaves an essentially random composite of σ^2 and σ_r^2, as in the analysis of the tree potentials in Table 7. For predictions from the average curve to the population of which the replicate equations are a sample, the error variances for a_o and for the regression coefficients a_1 and b_1 rarely contain quite the same components.

Error terms for each statistic

The error variance of each statistic, as derived by large sample theory, is in terms of the population variance σ^2, but in practice is solved with an estimated s^2 based upon the mean squares in an analysis of variance. The statistics a_o, a_1 and b_1 in Equation 3 or 6 have error variances similar to those for linear regression equations. The variance of a_o is

$$V(a_o) = \sigma^2/N \qquad (14)$$

for N values of the variate y, where our estimate of σ^2 is usually the mean square between replicates in an analysis of variance. In common with the linear regression coefficient, the error variance of a_1 and of b_1 is σ^2 divided by the denominator of the coefficient or

$$V(a_1) = V(b_1) = \sigma^2/f\Sigma u_1^2 = 2\sigma^2/fk \qquad (15)$$

where f is the number of replicates at each of k intervals in the cycle. The estimate of σ^2 will depend upon which of the variance components defined in Table 5 have proved significant in the analysis of variance.

The functions of a_1 and b_1 are of as much interest as the coefficients themselves. One of these, the semi-amplitude $A = \sqrt{a_1^2 + b_1^2}$, can be shown to have the same variance as the coefficients from which it is computed or

$$V(A) = \sigma^2/f\Sigma u_1^2 = 2\sigma^2/fk \qquad (16)$$

These variances are in units of y^2. In contrast, the variance of the phase angle $\theta = b_1/a_1$, is in terms of radians squared and is estimated as

$$V(\theta) = 2\sigma^2/fkA^2 \qquad (17)$$

This can be converted, of course, from radians to units of the original cycle. The square root of each variance is the standard error of its statistic. When computing confidence or fiducial limits for a given probability 1—P, the standard error is multiplied by the corresponding Student's t for the degrees of freedom n in the estimate of σ^2.

These estimates of precision may be illustrated with the example in Appendix Table 2 on the diurnal variation in the standing potential of an elm tree, which includes a trend. Since each variate y is the sum of the potentials at a given hour on three successive days, coded by changing the sign and subtracting 150, reversing the code and dividing by 3 converts each y to the original unit. Each mean square in Table 7 is decoded by dividing by 3^2. Because of the progressive decrease in the average potential through the period covered by the data, the estimate of a_0 and its error are contingent upon the date for which the equation is to be solved.

For any day (x) from August 1 to 25, 1953, inclusive, our estimate of the position of each curve is $a_0 = -60.359 - 0.4751\ x$. With this proviso, the variance of a_0 is computed with the mean square for the scatter about the trend, $947.9706/9 = 105.3301$ to obtain by Equation 14, $V(a_0) = 105.3301/192 = 0.54859$. At the mean date, $\bar{x}$ = August 13.25, the standard error of a_0 is $\sqrt{0.5486} = 0.7407$; at any other date its variance would be increased by the variance of the slope multiplied by $(x-\bar{x})^2$. Whenever the variation in T_r defines a trend, the estimate of a_0 is subject to a similar limitation.

Since the mean squares for both scatter and the interaction of replicate by (a_1+b_1) are here significant, the variance of the regression coefficients a_1 and b_1 is a linear combination of three mean squares, $s^2 = (15.6421 + 186.4238 - 5.3761)/9 = 21.8544$ with 15.50 degrees of freedom (Equation 12). The regression coefficients, $a_1 = 2.2017$ and $b_1 = 5.0279$, and the semi-amplitude, $A = \sqrt{30.1275} = 5.4889$, have identical variances: $V(a_1) = V(b_1) = V(A) = 2\times21.8544/8\times24 = 0.22765$, and a standard error of $\sqrt{0.22765} = 0.47713$.

The tangent of the phase angle θ can be computed without smoothing error from the coded numerators for a_1 and b_1 as $\tan \theta' = (-1448.035)/(-634.103) = 2.2836$. Since b_1 and a_1 are both positive after decoding, $\theta = \theta'$ and the phase angle is $\theta = 1.1580$ radians. Multiplying by $24/2\pi$ converts θ from radians to $24\times1.1580/6.2832 = 4.4234$ hours, as measured from our first reading at midnight, which places the maximum potential at 4:25 a.m. For the variance of θ, we have from Equation 17 and the variance of a_1, $V(\theta) = 0.22765/30.1275 = 0.007556$. In terms of

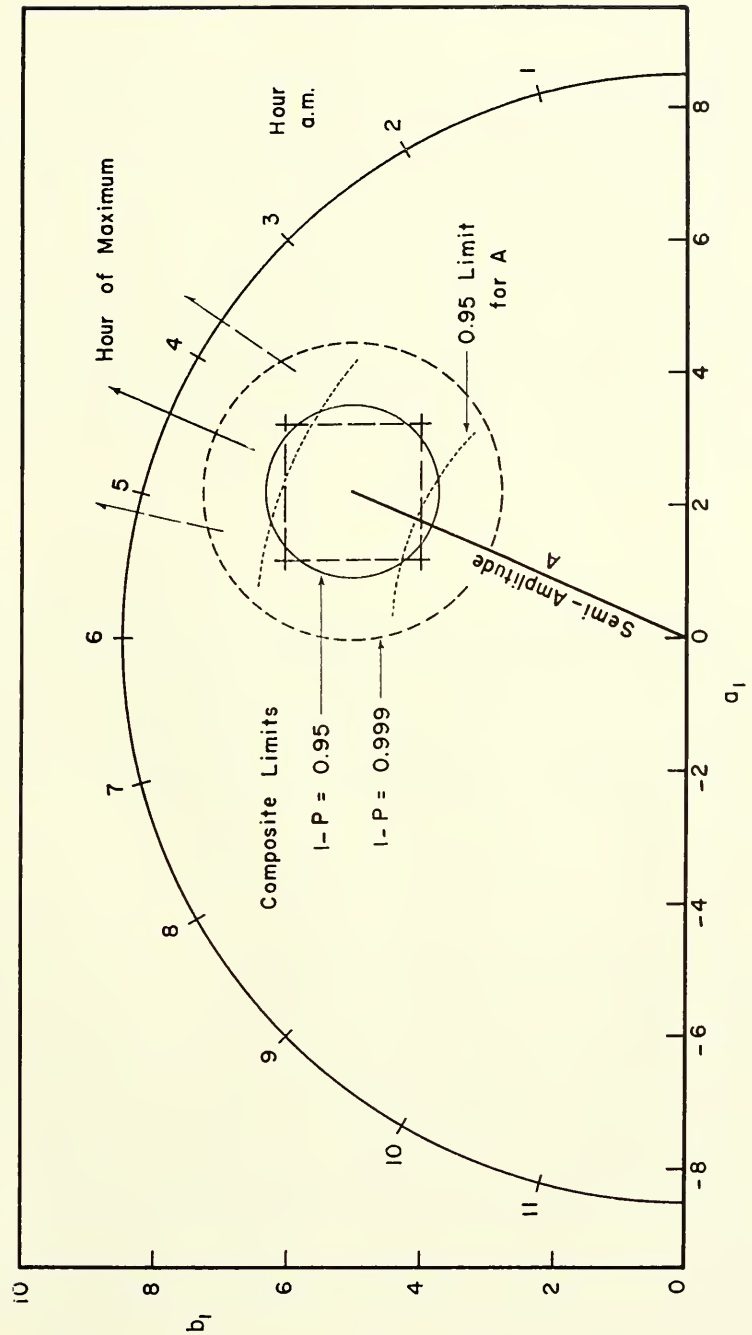

Figure 11. Confidence limits for the coefficients of the sine curve in Figure 5.

hours the phase angle has a standard error of $24 \times 0.08693/6.2832 = 0.3320$.

Each of these standard errors, with approximately 15.50 degrees of freedom, is multiplied by Student's $t = 2.1255$ at $P = 0.05$ in computing the 95% fiducial or confidence limits. For a_1, b_1 and A, the limits are $2.1255 \times 0.47713 = 1.0141$ above and below each statistic. Their relations are shown conveniently in Figure 11, where b_1 has been plotted on the ordinate against a_1 on the abscissa, and the clock hours are indicated on the half circle. When considered independently, the two regression coefficients are consistent at odds of 19 in 20 with any value of the parameter falling between the parallel horizontal or vertical lines bounding the point a_1,b_1. The corresponding interval for the semi-amplitude, the length of the solid line from zero to the point a_1,b_1, is defined by two parallel arcs with their centers at zero. The time of the maximum tree potential and its limits are marked by projections to the time scale on the half circle.

Composite tests

In estimating a separate interval for a_1 and for b_1, which would include its parameter in all but five percent of trials, we would reject their true values, considered jointly, with a frequency of $100(1 - 0.95^2) = 9.75$ percent. A more comprehensive approach is provided in Section 64 of "The Design of Experiments" by R. A. Fisher. If a_1 and b_1 are estimates of the true coefficients α_1 and β_1, the following inequality holds if the hypothesis is not to be contradicted at the percentage level selected for the variance ratio F:

$$(a_1 - \alpha_1)^2 + (b_1 - \beta_1)^2 \leqslant 2Fs^2/\tfrac{1}{2}kf \qquad (18)$$

where F is the tabular value with $n_1 = 2$ and $n_2 =$ the degrees of freedom in the relevant error variance s^2. The denominator converts the numerator, a sum of squares with two degrees of freedom, from units of a single variate y to that of the regression coefficients a_1 and b_1. Any pair of postulated regression coefficients α_1 and β_1 would be excluded if, in the quadratic form at the left, the differences were to exceed the limiting sum of squares on the right of the inequality.

All acceptable values of the parameters α_1 and β_1 then fall within a circle centered at the point a_1,b_1, which also defines the joint limits of the true amplitude and of the true phase angle. Its radius is the square root, of the right side of the above inequality or $\sqrt{2Fs^2/\tfrac{1}{2}kf}$. For the limits of the phase angle, the radius of the circle for any given probability is multiplied by $k/\pi A$ to convert it to the scale of k subdivisions in a complete cycle. The circle enclosing all acceptable parameter values at a selected level of significance may be drawn in a diagram not unlike Figure 11, and supple-

mented, if desired, with a series of concentric circles, one for each additional probability.

For our example on tree potentials, we may obtain indirectly from the table of z in Fisher and Yates (1957) $F = 3.6572$ for the 5% point and $F = 11.1471$ for the 0.1% point at $n_1 = 2$ and $n_2 = 15.50$. Substituting $F = 3.6572$ in Equation 18, any pair of postulated coefficients α_1 and β_1 which does not violate the inequality

$$(2.2017 - \alpha_1)^2 + (5.0279 - \beta_1)^2 \leqslant 1.6651$$

would be admitted at the 5% level by our observations. This pair of values would fall within a circle with a radius of $\sqrt{1.6651} = 1.2904$. Substituting F for the 0.1% level, we would have a larger concentric circle with a radius of 2.2528. These two circles have been added to Figures 11.

Finer Adjustments

Correction for length of month

In the annual cycles that we have been considering, the variate for each month has been given equal weight, although months differ in length by as much as 10%. The month containing the maximum or minimum variate has been estimated with an "average" month of $1461/48 = 30.4375$ days, and the date within the selected month then based upon its length. In a paper of the Meteorological Research Committee (London), Craddock (1955) has provided an adjusted set of multipliers which allows for differences in the length of the month. With these multipliers, the coefficients for a two-term harmonic equation can be computed as readily as with the orthogonal cosines and sines in Table 1. For computing the corrected expectations Y_c, he provides a second table of the cosines and sines for each month. Since it is not orthogonal, his equation cannot be reduced immediately from two terms to one term or extended to a third or higher term as the data require. For describing the annual course of the mean temperature in the northern hemisphere, this limitation is negligible, since Craddock has found that a two-term Fourier series applies quite generally.

As an indication of the size of the correction with relatively precise data, the monthly mean temperatures in Table 2 have been fitted by both methods. When computed from the totals T_t by Equation 9, weighting each month equally and with exact values for Σu_i^2 and Σv_i^2 instead of their expectations, $\frac{1}{2}k = 6$, we have the two-term harmonic series:

$$Y = 50.7655 + 20.9898a_1 + 3.4853b_1 - 0.1060a_2 + 0.6825b_2$$

starting with July as t_0. The monthly means for this 14-year period $\bar{y}$ and their expectations Y from the above equations are listed in Table 13. When

TABLE 13. Comparisons of the observed monthly mean temperatures in New Haven for 14 years ($\bar{y}$) and their predicted values from two-term Fourier equations computed without weighting (Y), with corrections for the length of each month (Y_c), and with the weights w (Y_w) in Table 14.

Month	Observed	Unweighted	Difference between means		
	$\bar{y}$	Y	$\bar{y}-Y$	Y_c-Y	Y_w-Y
Jul	72.4357	71.6494	.7863	−.0071	.1765
Aug	70.9286	71.2234	−.2948	.0017	.0077
Sep	64.1357	64.9228	−.7871	.0185	−.1965
Oct	54.8714	54.3568	.5146	.0067	−.2137
Nov	43.4714	42.7508	.7206	−.0265	−.0064
Dec	32.6929	33.6869	−.9940	−.0387	.2358
Jan	29.9214	29.6697	.2517	−.0059	.2806
Feb	31.4571	31.3837	.0734	.0370	.0828
Mar	37.9214	37.8963	.0251	−.0752	−.1702
Apr	47.8000	47.3861	.4139	−.0325	−.2435
May	56.8714	57.7040	−.8326	−.0081	−.0841
Jun	66.6786	66.5560	.1226	−.0054	.1308

$\Sigma(\bar{y}-Y)^2 = 4.04575$, $\Sigma(Y_c-Y)^2 = 0.01085$, $\Sigma(Y_w-Y)^2 = 0.36917$

corrected for differences in the length of the successive months but starting in January as t_0, we have with Craddock's weighted multipliers the two-term harmonic equation:

$$Y_c = 50.8623 - 19.7304a_1 - 8.6060b_1 - 0.4865a_2 + 0.5148b_2$$

The corrected predictions for each month Y_c were then computed with Craddock's parallel table of cosines and sines and the constant $a_0 = 50.8623$.

The discrepancies (Y_c-Y) may be compared with the deviations $(\bar{y}-Y)$ of the observed means from their simpler predictions Y. They are clearly of a different order of magnitude. Comparing their sums of squares, $100\Sigma(Y_c-Y)^2/\Sigma(\bar{y}-Y)^2 = 0.27$ percent. If this single example can be considered a reliable indicator, the discrepancy due to computing the Fourier regression coefficients as if months were equal in length should be negligible for most purposes.

Variance homogeneity

A second discrepancy between theory and observation may be traced to our assumption of equal variability at successive intervals through the cycle. Climatologists, for example, have long known that the variation in temperature from year to year in a given locality is greater in winter than in summer. To the extent that this inequality represents harmonic variation, either in amplitude or in phase, it should be attributable to differences between the curves fitted separately to the data for each year. If this explanation were fully effective, the deviations of the observed monthly means from the fitted annual curves should be of the same magnitude in each month through the year. The problem is important in predicting the size of discrepancies from the fitted curve, and in determining the best estimate of the mean curve over the several replicates.

When comparing the observed temperature in each interval with its expectation, approximations in curve fitting which are entirely edequate in an overall analysis may prove troublesome. Sums of the squared individual deviations may differ from their counterparts in the analysis of variance in the third and even in the second significant figure due to apparently negligible rounding errors, especially if the average Fourier curve absorbs a very large proportion of the total sum of squares. As in the calculation of a reciprocal matrix, a good numerical check may depend upon carrying what seems initially to be an unreasonable number of decimal places. An example is our substitution of the true value $\frac{1}{2}k$ for Σu_i^2 and Σv_i^2 in the denominator of the Fourier coefficients, some of which are irrational numbers rounded to three decimal places. In a cycle of twelve subdivisions, this substitutes $\frac{1}{2}k = 6$ for $\Sigma u_1^2 = \Sigma v_1^2 = 5.999824$, $\Sigma u_2^2 = 6$ exactly and $\Sigma v_2^2 = 5.999648$, the sums of squares of the rounded coefficients. These latter values have been used in the following analysis.

Because the second term in the Fourier series has varied significantly from year to year, it has been retained in a closer analysis of the monthly mean temperatures in New Haven in Table 2. As a first step, a separate two-term Fourier equation (Equation 9) has been computed from the 12 monthly means (y) for each year. Each of these 14 equations was then solved 12 times, with the u_i and v_i for $t = 0$ to 11, leading to a table of predicted means, designated here as $\hat{y}$, which parallel the y's in Table 2. The averages of the 14 $\hat{y}$'s, one for each month, agreed exactly with the Y's in Table 13 computed independently with the two-term equation based upon the monthly totals of the y's, T_t. Each $\hat{y}$ was then subtracted from its corresponding observed mean temperature y in Table 2, to obtain the deviations $d = (y - \hat{y})$ in Appendix Table 7. These total zero for each year, and for each month their average is equal to the difference $(\bar{y} - Y)$ in Table 13.

TABLE 14. Observed monthly variances (per degree of freedom) of New Haven mean temperatures for $(\bar{y}-Y)^2$ from the observed means $\bar{y}$ and their unweighted predictions Y in Table 13, V(y) from the deviations of the y's in Table 2 from their column means $\bar{y}$, $V(\hat{y})$ from the deviations $(\hat{y}-\bar{y})$, and V(d) from the deviations $(y-\hat{y})$ in Appendix Table 7; expected standard deviations (SD) from the sine curve fitted to log-V(y); weights w = antilog(1 − log-V(d)) for computing the weighted two-term Fourier curve Y_w in Table 13.

Month	Observed variance from				SD from	w
	$(\bar{y}-Y)^2$	V(y)	$V(\hat{y})$	V(d)	log-$\hat{V}$(y)	
Jul	14.840	3.538	7.348	2.039	1.683	5.4
Aug	2.087	3.401	7.475	1.309	1.680	4.8
Sep	14.866	2.904	4.162	3.206	1.853	3.5
Oct	6.357	4.564	4.018	5.123	2.197	2.4
Nov	12.465	4.851	10.678	4.805	2.679	1.6
Dec	23.715	12.170	20.887	8.533	3.183	1.2
Jan	1.521	19.450	21.260	12.752	3.519	1.1
Feb	.129	11.101	13.342	6.257	3.525	1.3
Mar	.015	9.560	11.171	5.670	3.198	1.7
Apr	4.112	5.832	11.388	2.432	2.696	2.6
May	16.636	4.798	6.577	4.291	2.212	3.8
Jun	.361	3.440	4.537	2.275	1.861	4.9
Mean	8.092	7.134	10.275	4.891	2.434	34.3
Σ(DF)	7	156	65	91		= T

Four variances were then determined for each month in units of the variance of a single monthly temperature y. The average of each series of variances over the 12 months agreed with its corresponding mean square from the analysis of variance, in several cases combining sums of squares that were reported initially in separate rows. The series of variances in Table 14 have the following composition:

Those from $(\bar{y}-Y)^2$ measure the discrepancy of the observed 14-year average for each month from that computed with the two-term Fourier equation for all 14 years, each with 7/12 of a degree of freedom. These deviations would be absorbed completely by the remaining terms of the Fourier series if it were extended to the limit.

The empirical variances $V(y) = \Sigma(y-\bar{y})^2/13$ represent the variation of the 14 y's for each month in Table 2 about their observed or column mean $\bar{y}$. They show a marked seasonal trend. Each sum of squares $\Sigma(y-\bar{y})^2$ with 13 degrees of freedom has been divided into two parts to obtain the next two series of variances.

The variances $V(\hat{y})$ measure the variation of the predicted $\hat{y}$'s about their mean, or that part of the variation in each month which is attributable to the 14 annual two-term Fourier curves. The average of the $V(\hat{y})$'s with 65 degrees of freedom is equal to the mean of the sums of squares in rows 1+5+6 of the analysis of variance (Table 6). These monthly variances, each with $65/12 = 5.4167$ degrees of freedom, absorb part, at least, of the seasonal trend in the variance.

The variances V(d), averaging less than half of the $V(\hat{y})$'s, represent our nearest approach to a random error. They have been computed from the differences d for each month in Appendix Table 7 as $V(d) = 12\Sigma(d-\bar{d})^2/91$, each with 91/12 degrees of freedom. Their mean corresponds in the analysis of variance to the mean square in row 7. Although much of the initial seasonal trend in the variance has been absorbed by the $V(\hat{y})$'s, a substantial amount still persists.

The pattern of the seasonal trend in the empirical variance V(y) and in its two components in Table 14 may be defined periodically. Since the distribution of the log-variance is approximately normal (Bartlett, 1947), the following sine curves have been fitted to their logarithms and plotted in Fig. 12:

$$\text{Log-}\hat{V}(y) = 0.7727 - 0.3203u_1 - 0.0888v_1 \quad (s^2 = 0.01198)$$

$$\text{Log-}\hat{V}(\hat{y}) = 0.9490 - 0.2585u_1 - 0.0662v_1 \quad (s^2 = 0.02532)$$

$$\text{Log-}\hat{V}(d) = 0.6071 - 0.3384u_1 + 0.0165v_1 \quad (s^2 = 0.02106)$$

In no case was the second Fourier term significant. By Equation 4, the semi-amplitudes (A) of these curves are respectively 0.3324 ± 0.0119, 0.2668 ± 0.0174, and 0.3392 ± 0.0158. From antilog (2A) for each series, the smallest expected variance in the mean summer temperature would be multiplied by a factor of 4.62 for y, 3.42 for $\hat{y}$, and 4.77 for d to abtain the largest winter variance. From the phase angle for each curve, the variances were maximal on January 31, 30 and 12 respectively.

Weighted periodic curves

The variance of the mean temperature differs sufficiently through the year from the equality implied in our initial model, that a weighted analysis might be expected to improve our estimate. Appropriate weights would be the reciprocals of the expected random variance, computed from the sine curve for log-V(d) as w = antilogarithm of 1 — log-V(d). These weights, in the last column of Table 14, vary from 1.1 to 5.4 and resemble the second of the three weighting systems suggested by Craddock (1955)

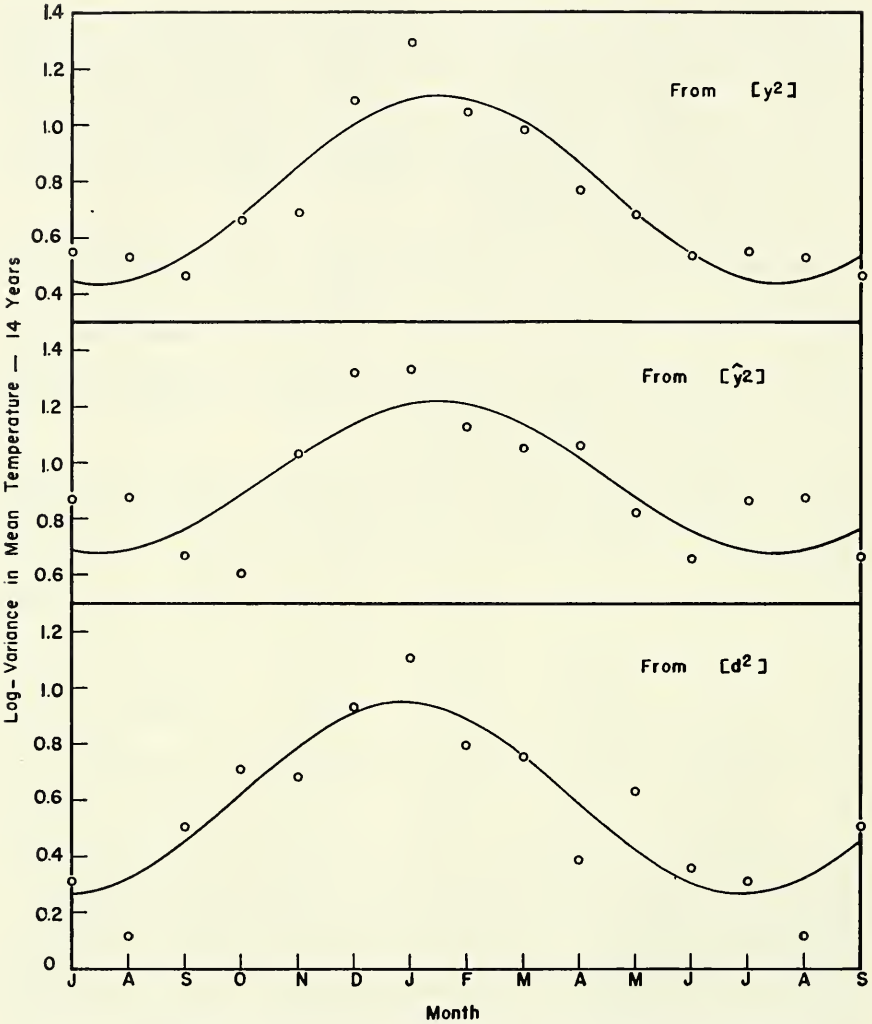

Figure 12. Seasonal variation in the logarithm of the variances in Table 14, from $[y^2] = \Sigma(y-\bar{y})^2$ for the overall deviations in the monthly mean temperatures, and from its components $[\hat{y}^2] = \Sigma(\hat{y}-\bar{y})^2$ and $[d^2] = \Sigma(y-\hat{y})^2$, each fitted with a sine curve.

for a similar purpose. The weighted two-term Fourier curve, computed by partial regression, has the equation:

$$Y_w = 50.7655 + 20.9378u_1 + 3.5002v_1 + 0.1226u_2 + 0.6028v_2$$

When solved with the cosines and sines for each month, the weighted mean temperatures Y_w differ from the unweighted expected means Y as shown by the differences $(Y_w - Y)$ in the last column of Table 13.

The sum of squares from these differences is a considerably larger fraction (9.12%) of $\Sigma(\bar{y}-Y)^2$ than the 0.27 percent for the corresponding differences (Y_c-Y). Although the weighted estimates Y_w may be superior theoretically, their curve requires the solution of a reciprocal matrix and gives considerably more weight to the summer than to the winter temperatures. From a commonsense point of view, one may question whether the weighted estimates are as satisfactory climatologically as those from the unweighted Fourier equation, to which each month contributes equally. Is it wise to base the estimate of the annual curve so largely upon the summer months?

Normality of temperature deviations

In analyses of variance of periodic regressions we tacitly assume not only that the random deviations are equally variable at each t but also that their distribution is normal. Because of the small number of years in our climatological example, the normality of the deviations d has been tested graphically. The rankits* for a sample of 14 have been plotted in Figure 13 against the deviations for each month in Appendix Table 7 in rank order and each fitted with a straight line. Their slopes are less in winter than in summer, as would be expected from the seasonal change in the variance. If the distributions are normal, the plotted points should not curve systematically from the computed straight lines. To test whether the trends in Figure 13 cancel out, the deviations may be averaged for each position over the 12 months (i.e., the largest in each month, the next largest, etc). The rankits have been plotted against these averages in the left side of Figure 14 and fitted with a line passing through 0,0 with a slope of $1/s = 0.5920$, where $s = \sqrt{445.0708/12\times13} = 1.6891$. The close agreement with a straight line confirms our initial hypothesis that the variation about the two-term Fourier series for each year is here essentially normal.

Two aspects of periodic regression need to be distinguished. The first is the harmonic analysis of periodic data to determine their underlying pattern and the magnitude and nature of the variation to which this pattern has been exposed. The second problem is that of predicting future responses from our present data, as is commonly the objective in climatology. Unless the constants in our fitted Fourier curve for each year were to define a trend which might be expected to continue, and climatologists are not agreed upon the existence of these trends, the prediction of future temperatures would have to be based upon the average curve for past years.

* A rankit is the expected mean deviation for each rank in an ordered sample of a given size from a normal population with a mean of zero and a standard deviation of one (Fisher and Yates, 1957, Table XX).

The error in our prediction would then involve not only the variation around the annual curves, which seems to be satisfactorily normal although not constant, but also the variation of the annual curves about their average for the series of years. When these two sources of variation are combined, a convenient estimate of the standard deviation for each month in

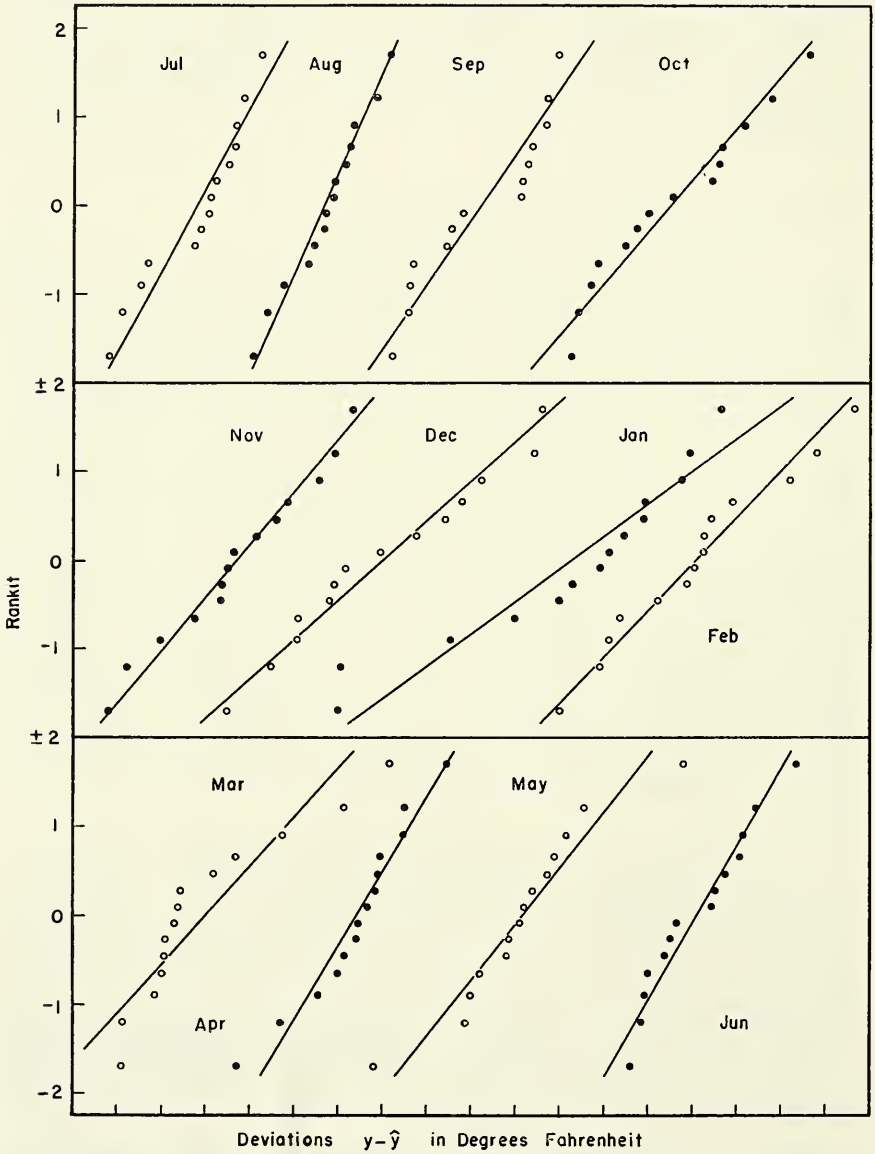

Figure 13. Rankit test for agreement of the deviations in the monthly mean temperature with the normal distribution, from the differences d = (y−ŷ) in Appendix Table 7.

°F is SD = antilogarithm of $\frac{1}{2}(\log\text{-}\hat{V}(y))$ in Table 14, from the equation of the upper sine curve in Figure 12. There is no assurance *a priori*, however, that the composite variation will be distributed normally.

For a graphic test of normality, the deviations $(y-Y)$, which also include the differences $(\bar{y}-Y)$, have been computed from the y's in Table 2 and the Y's in Table 13. These were ranked in order for each month and then averaged over the twelve months to obtain the rankit diagram in the right side of Figure 14. The plotted points have been fitted with a straight line passing through 0,0 with a slope of $1/s = 1/2.679$. Not only is the slope much less than that for the deviations about the annual fitted curves in the left side of the figure, but the points themselves describe a trend that is less certainly linear.

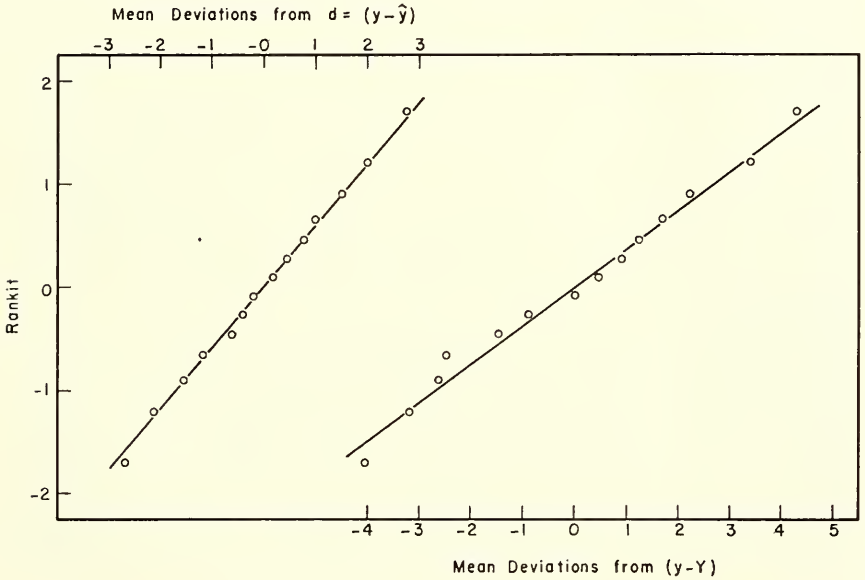

Figure 14. Test for normality of the ranked deviations $(y-\hat{y})$ in Figure 13 average over the 12 months (left curve), compared with a similar diagram of the average deviations $(y-\bar{y})$ from the 14-year means for each month $(\bar{y})$.

Despite their limited sensitivity with as few as 14 replicates, the numerical measures of skewness (g_1) and kurtosis (g_2) have the advantage of separating these two types of non-normality (Fisher, 1954). Both statistics are normally distributed about zero with a standard error depending only upon the size of the sample. They have been computed for each month from the distribution of the observed temperatures y about their monthly means $\bar{y}$; neither approached significance in any one month or in composite χ^2 tests over all 12 months. On the off chance that a seasonal trend might still be discernable, separate sine curves have been fitted to the

twelve monthly values for g_1 and for g_2. Neither periodic trend approached significance ($P > 0.20$), but their minima and amplitudes may be suggestive. The curve for g_1 had a minimum in December and an amplitude of 0.507 ± 0.338, and that for g_2 a minimum in January and an amplitude of 1.019 ± 0.552. In developing probability statements for the monthly mean temperature from more extensive data, we may need to consider not only seasonal changes in the standard deviation about the average two-term Fourier curve, but also seasonal departures from normality.

Summary

Periodic regression is applied here to cyclic phenomena in biology and climatology in which (1) the length of the cycle, such as a year or day, is determined independently of the response, (2) the observations are spaced evenly through the cycle, and (3) the number of replicates is constant at each interval. When the response (y) changes symmetrically through the cycle, the first harmonic or sine curve is defined by the mean response (a_0) and two orthogonal regression coefficients, a_1 for the cosine u_1 and b_1 for the sine v_1, as $Y = a_0 + a_1u_1 + b_1v_1$, from which we can compute its amplitude and phase angle. When the curve is not symmetrical, the sine curve can be extended with additional terms for two, three or more cycles in each fundamental period by classical Fourier analysis until the desired fit is achieved.

The analysis of variance for deciding how many terms to retain in a Fourier curve and for determining its error is based upon the mathematical model for replicated regressions. In effect, a Fourier curve is fitted to each replicate and the analysis determines in what respects these separate curves differ from replicate to replicate. Various aspects of the calculation are illustrated by the monthly mean temperatures in New Haven over a 14-year period, the monthly iodine values of butterfat from five creameries in Alberta, and the electrical potential of an elm tree in eight three-day periods in August, 1953.

Both the number of terms in a periodic regression and the validity of its analysis depend upon the selection of a suitable unit for the response. The transformation to logarithms is applied to monthly data on two contagious diseases. The square root transformation for counts is illustrated with data on the hour of birth, where agreement with the assumed Poisson variation about a diurnal sine curve can be tested by χ^2. The analysis of seasonal variation in the log-ED50 for a biocide or a drug is computed by maximum likelihood from all-or-none data with probits. A periodic response can be corrected for differences in a concomitant environmental factor by covariance, as illustrated by the adjustment for aperiodic humidity of the diurnal variation in the log-heat exchange of cows in an experimental barn.

The precision of the constants for the first harmonic in a periodic regression is considered from two viewpoints. The first defines the variance of the statistics of a sine curve and the confidence limits of their parameters when each statistic is considered separately. The second defines a joint circular region within which any combination of the parameters for a_1 and b_1 is compatible with our observations at a given probability.

Finer adjustments in periodic regression are examined with the monthly mean temperatures in New Haven. A correction for differences in the length of the month proved of minor importance relative to other errors in fitting a two-term Fourier curve. Seasonal changes in the variance through the year could be divided into two components, one representing differences between the observed monthly temperatures and their predictions by annual two-term Fourier curves, and the other differences between these predicted temperatures and the average two-term Fourier curve for all 14 years. For each source the log-variance changed periodically through the year in a sine curve, leading to estimated standard deviations for probability predictions and to weights for recomputing the average Fourier curve.

A distinction is drawn between two objectives in periodic analysis, that of locating sources of variation and describing their characteristics, and forecasting, which must ordinarily be based upon the average over all replicates because of the unpredictable nature of long term trends. The summer temperatures contributed proportionately more to the weighted regression than the winter temperatures, a feature which may be potentially less desirable for climatological predictions than the simpler process of equal weighting through the year. In graphic tests with rankits, the approximately random deviations from the yearly two-term Fourier curves proved to be satisfactorily normal but when these were increased by the larger differences between the yearly and the average curves, the data suggest that seasonal departures from normality may modify probability predictions based upon an average Fourier curve.

Acknowledgements

For his personal advice and invaluable aid on measuring the precision of periodic regressions, I have Professor Sir Ronald Fisher to thank. For generously supplying me with original data, I am indebted to Professor H. S. Burr (Appendix Table 2), to Dr. P. T. King (Appendix Table 5), and to Dr. R. E. Serfling (Appendix Table 4). Christopher Bingham, T. W. Gamelin and Agnes McNamara have assisted most helpfully with the calculations. The statistical development and presentation have benefited from the advice of Professors R. E. Bargmann, Jerome Cornfield, and T. C. Koopmans, Dr. H. C. S. Thom, and my colleagues at this Experiment Station, especially Dr. P. E. Waggoner. C. I. B.

References

Aitken, A.C. (1939) Statistical Mathematics. Oliver and Boyd, Edinburgh.

Anderson, R.L. and Anderson, T.W. (1950) Distribution of the circular serial correlation coefficient for residuals from a fitted Fourier series. Ann. Math. Stat. *21*:59-81.

Anderson, R.L. and Bancroft, T.A. (1952) Statistical Theory in Research. (p. 350) McGraw-Hill Book Co., Inc., New York.

Bartlett, M.S. (1936) The square-root transformation in analysis of variance. J. Roy. Stat. Soc. Suppl. *3*:68-78.

Bartlett, M.S. (1947) The use of transformations. Biometrics *3*:39-52.

Bliss, C.I. (1952) The Statistics of Bioassay, with Special Reference to the Vitamins. Academic Press, N.Y.

Brooks, C.E.P. and Carruthers, N. (1953) Handbook of Statistical Methods in Meteorology. Her Majesty's Stationery Office, London.

Burr, H.S. (1945) Diurnal potentials in the maple tree. Yale J. Biol. and Med. *17*:727-734.

Craddock, J.M. (1955) The variation of the normal air temperature over the Northern Hemisphere during the year. Meteorological Research Committee (London) M.R.P. No. 917.

Finney, D.J. (1952) Probit Analysis. 2nd Ed. University Press, Cambridge, England.

Fisher, R.A. (1951) The Design of Experiments. 6th Edition. Oliver and Boyd, Edinburgh.

Fisher, R.A. (1954) Statistical Methods for Research Workers. 12th Edition. Oliver and Boyd, Edinburgh.

Fisher, R.A. and Yates, F. (1957) Statistical Tables for Biological, Agricultural and Medical Research. 5th Edition. Oliver and Boyd, Edinburgh.

Kendall, M.G. (1945) The Advanced Theory of Statistics. 2nd Edition. Volume 1, section 5·10. Chas Griffin and Co., Ltd., London.

King, P.D. (1956) Increased frequency of births in the morning hours. Science *123*:985-986.

Metropolitan Life Insurance Co. (1945-55) Mortality from selected causes. Statistical Bull. Vols. 26-36. New York.

Penhos, J.C., Cornfield, J. and Cordero Funes, J.R. (1954) Variacion estacional de la espermiacion del sapo por la gonadotrofina corionica. Revista Sociedad Argentina de Biologia *30*:77-87.

Serfling, R.E. and Sherman, I.L. (1953) Poliomyelitis distribution in the United States. Public Health Reports *68*: 453-466.

Thompson, H.J. (1954) Environmental physiology and shelter engineering. XXIV. Effect of temperature upon heat exchanges in dairy barns. Mo. Agr. Expt. Sta. Research Bull. 542.

U.S. Department of Commerce, Weather Bureau (1956-57) Local climatological data with comparative data, New Haven, Conn.

Wood, F.W. (1956) Seasonal and regional variations in some chemical and physical properties of Alberta butterfat. Can. J. Agric. Science *36*:422-429.

Added in proof: When the length of the primary cycle must be estimated from the data, an approach related to that of the present bulletin is given by H. O. Hartley (1949) Tests of significance in harmonic analysis. Biometrika *36*:194-201.

APPENDIX TABLE 1. Average monthly iodine value (−33.0) of butterfat from five creameries in Alberta, Canada, each based on 104 consecutive weekly samples beginning in April 1952. (Wood, 1956)

Locality	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	T_r
Edmonton B	4.6	6.9	6.3	7.0	8.4	8.4	5.1	2.7	2.9	3.5	4.0	4.3	64.1
Edmonton A	4.9	6.9	6.1	6.5	7.2	7.5	4.3	2.6	2.5	2.9	4.0	3.9	59.3
Red Deer	4.2	5.5	4.4	5.5	6.2	5.8	3.0	1.5	2.4	2.8	2.4	2.5	46.2
Coronation	4.2	4.9	3.3	3.3	4.3	4.0	2.6	.9	.7	1.8	2.5	2.1	34.6
Calgary	4.7	4.9	2.8	3.0	3.2	3.2	1.4	.5	1.6	2.2	3.1	2.5	33.1
T_t	22.6	29.1	22.9	25.3	29.3	28.9	16.4	8.2	10.1	13.2	16.0	15.3	237.3
Mean $\bar{y}_t$	4.52	5.82	4.58	5.06	5.86	5.78	3.28	1.64	2.02	2.64	3.20	3.06	3.955
Predicted Y	4.66	5.48	4.92	4.87	5.92	5.73	3.39	1.54	1.96	2.90	2.88	3.21	3.955
From (a_1+b_1)	.409	1.220	1.704	1.732	1.295	.512	−.409	−1.220	−1.704	−1.732	−1.295	−.512	.000
" (a_2+b_2)	.070	−.445	−.515	−.070	.445	.515	.070	−.445	−.515	−.070	.445	.515	.000
" (a_3+b_3)	.223	.747	−.223	−.747	.223	.747	−.223	−.747	.223	.747	−.223	−.747	.000

APPENDIX TABLE 2. Hourly standing potentials in an elm tree in Lyme, Connecticut, in August, 1953. (H. S. Burr, 1958)

Hour	y = −Σ (Daily potential) — 150 on August								Total T_t
	1-3	4-6	8-10	11-13	14-16	17-19	20-22	23-25	
Mt.	23	30	41	38	41	43	50	76	342
1	23	29	39	33	41	42	49	71	327
2	23	24	32	32	36	41	48	68	304
3	20	22	36	31	34	39	46	65	293
4	20	22	36	28	33	38	44	58	279
5	19	22	30	30	32	41	43	57	274
6	20	22	30	28	32	42	42	57	273
7	20	22	32	30	33	42	41	57	277
8	17	31	36	37	37	44	39	57	298
9	20	39	40	49	38	41	41	59	327
10	22	51	44	58	41	50	46	65	377
11	26	53	55	64	52	54	57	75	436
12	32	57	64	69	62	56	65	79	484
1	32	60	69	70	67	58	72	79	507
2	32	63	70	70	63	61	74	79	512
3	35	63	70	70	62	65	74	81	520
4	38	58	70	70	61	66	76	80	519
5	38	58	68	70	60	68	77	80	519
6	38	63	64	70	57	68	77	78	515
7	40	63	60	67	57	67	73	78	505
8	34	57	57	56	53	62	68	78	465
9	35	55	54	51	50	56	64	79	444
10	23	54	51	49	43	49	58	80	407
11	20	54	47	48	41	45	56	78	389
T_r	650	1072	1195	1218	1126	1238	1380	1714	9593

APPENDIX TABLE 3. Deaths from pneumonia for ten years starting September 1945, in terms of y = log(annual rate per 100,000) −0.800. (Metropolitan Life Insurance Co., 1945-55)

Year	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	T_r	x_1	x_2
1945-46	.409	.462	.501	.666	.869	.769	.749	.655	.486	.534	.464	.364	6.928	−9	6
1946-47	.372	.484	.503	.569	.673	.603	.658	.727	.557	.412	.372	.270	6.200	−7	2
1947-48	.335	.361	.439	.542	.699	.628	.676	.566	.555	.446	.344	.300	5.891	−5	−1
1948-49	.335	.304	.367	.409	.588	.555	.548	.571	.501	.313	.270	.332	5.093	−3	−3
1949-50	.290	.316	.355	.422	.481	.448	.528	.552	.474	.358	.200	.140	4.564	−1	−4
1950-51	.144	.264	.245	.396	.443	.467	.673	.565	.446	.276	.245	.191	4.355	1	−4
1951-52	.057	.237	.321	.430	.546	.443	.601	.499	.428	.297	.233	.168	4.260	3	−3
1952-53	.124	.225	.358	.382	.467	.631	.635	.479	.314	.297	.286	.114	4.312	5	−1
1953-54	.279	.154	.200	.390	.393	.415	.446	.379	.279	.264	.221	.221	3.641	7	2
1954-55	.209	.237	.237	.334	.483	.472	.455	.412	.249	.204	.268	.245	3.805	9	6
T_t	2.554	3.044	3.526	4.540	5.642	5.431	5.969	5.405	4.289	3.401	2.903	2.345	49.049	0	0

APPENDIX TABLE 4. Monthly log-incidence of poliomyelitis in the United States in cases per 1,000,000 population, from April 1942 to March 1957. (Serfling and Sherman, 1958)

Year	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Total T_r
42-43	.675	.848	1.038	1.531	1.836	1.940	1.830	1.514	1.336	1.111	.937	.944	15.540
43-44	.856	1.077	1.628	2.114	2.407	2.538	2.236	1.905	1.510	1.016	.956	.818	19.061
44-45	.907	1.062	1.575	2.293	2.679	2.695	2.407	2.033	1.600	1.020	1.182	1.047	20.500
45-46	1.051	1.200	1.671	2.048	2.467	2.530	2.327	1.992	1.666	1.235	1.100	1.098	20.385
46-47	1.063	1.386	1.901	2.407	2.814	2.723	2.533	2.173	1.794	1.388	1.211	1.081	22.474
47-48	1.015	1.116	1.338	1.749	2.208	2.483	2.211	1.905	1.521	1.152	.996	.993	18.687
48-49	1.058	1.645	1.964	2.469	2.668	2.798	2.564	2.382	2.062	1.586	1.348	1.302	23.846
49-50	1.214	1.517	2.100	2.622	3.061	2.910	2.614	2.375	1.995	1.573	1.534	1.458	24.973
50-51	1.315	1.598	1.961	2.380	2.739	2.824	2.697	2.492	2.055	1.698	1.471	1.251	24.481
51-52	1.308	1.430	1.887	2.375	2.756	2.732	2.529	2.274	2.058	1.631	1.505	1.342	23.827
52-53	1.388	1.635	2.117	2.716	3.057	3.107	2.882	2.445	2.144	1.738	1.490	1.330	26.049
53-54	1.459	1.724	2.123	2.618	2.834	2.814	2.531	2.218	2.019	1.654	1.564	1.421	24.979
54-55	1.495	1.766	2.124	2.578	2.797	2.867	2.684	2.306	1.922	1.491	1.301	1.293	24.624
55-56	1.502	1.798	1.963	2.393	2.807	2.712	2.404	2.112	1.769	1.454	1.337	1.332	23.583
56-57	1.335	1.494	1.828	2.205	2.460	2.375	2.043	1.769	1.529	1.205	1.099	.960	20.302
T_t	17.641	21.296	27.218	34.498	39.590	40.048	36.492	31.895	26.980	20.952	19.031	17.670	333.311

APPENDIX TABLE 5. Number of normal human births in each hour in four hospital series, transformed to y = $\sqrt{\text{No. of births}}$. (King, 1956)

Hour starting		$\sqrt{\text{births}} = y$ in Hospital A	B	C	D	Total	Observed $\bar{y}$	Expected Y
Mt	12	13.56	19.24	20.52	21.14	74.46	18.6150	18.463
AM	1	14.39	18.68	20.37	21.14	74.58	18.6450	18.812
	2	14.63	18.89	20.83	21.79	76.14	19.0350	19.129
	3	14.97	20.27	21.14	22.54	78.92	19.7300	19.393
	4	15.13	20.54	20.98	21.66	78.31	19.5775	19.587
	5	14.25	21.38	21.77	22.32	79.72	19.9300	19.697
	6	14.14	20.37	20.66	22.47	77.64	19.4100	19.716
	7	13.71	19.95	21.17	20.88	75.71	18.9275	19.641
	8	14.93	20.62	21.21	22.14	78.90	19.7250	19.479
	9	14.21	20.86	21.68	21.86	78.61	19.6525	19.240
	10	13.89	20.15	20.37	22.38	76.79	19.1975	18.941
	11	13.60	19.54	20.49	20.71	74.34	18.5850	18.602
M	12	12.81	19.52	19.70	20.54	72.57	18.1425	18.246
PM	1	13.27	18.89	18.36	20.66	71.18	17.7950	17.897
	2	13.15	18.41	18.87	20.32	70.75	17.6875	17.579
	3	12.29	17.55	17.32	19.36	66.52	16.6300	17.315
	4	12.92	18.84	18.79	20.02	70.57	17.6425	17.121
	5	13.64	17.18	18.55	18.84	68.21	17.0525	17.011
	6	13.04	17.20	18.19	20.40	68.83	17.2075	16.993
	7	13.00	17.09	17.38	18.44	65.91	16.4775	17.067
	8	12.77	18.19	18.41	20.83	70.20	17.5500	17.229
	9	12.37	18.41	19.10	21.00	70.88	17.7200	17.468
	10	13.45	17.58	19.49	19.57	70.09	17.5225	17.767
	11	13.53	18.19	19.10	21.35	72.17	18.0425	18.106
Total		327.65	457.54	474.45	502.36	1762.00	18.3542	
$\Sigma (u_i y)$		3.25792	−3.42395	2.77825	2.59608	5.20830	.10851	
$\Sigma (v_i y)$		10.62339	19.04199	20.38840	15.29826	65.35204	1.36150	

APPENDIX TABLE 6. Hourly humidity or mixing ratio, $x = 2(H_2O/\text{dry air}) - 0.8$, and average heat exchange per cow, $y = \log (BTU/10^3)$, in an experimental dairy barn on 3 days in 1949. (Thompson, 1954)

Hour	Mixing ratio, x on				Log-BTU, y on				Adj.
starts	10/11	10/30	11/20	T_t	10/11	10/30	11/20	T_t	$\bar{y}$
3 pm	1.3	.5	.6	2.4	.512	.407	.423	1.342	.4310
4	1.0	.5	.5	2.0	.484	.415	.447	1.346	.4471
5	1.1	.7	.4	2.2	.550	.512	.462	1.524	.4991
6	.9	.8	.4	2.1	.512	.512	.477	1.501	.4951
7	.8	.7	.6	2.0	.505	.512	.512	1.529	.5044
8	1.0	.4	.6	2.0	.613	.462	.532	1.607	.5341
9	.9	.6	.7	2.2	.607	.498	.525	1.630	.5344
10	1.1	.6	.7	2.4	.623	.505	.505	1.633	.5280
11	.8	.5	.7	2.0	.525	.498	.519	1.542	.5124
12	.9	.4	.5	1.8	.538	.470	.491	1.499	.5055
1 am	1.0	.3	.3	1.6	.550	.470	.484	1.504	.5145
2	.4	.4	.6	1.4	.519	.477	.477	1.473	.5116
3	.4	.2	.3	.9	.532	.431	.498	1.461	.5260
4	.6	.7	.7	2.0	.371	.407	.447	1.225	.4068
5	1.0	.9	.9	2.8	.447	.491	.491	1.429	.4452
6	.9	.5	.7	2.1	.439	.431	.462	1.332	.4387
7	1.2	.7	.5	2.4	.505	.477	.470	1.452	.4677
8	.7	.8	.8	2.3	.415	.477	.491	1.383	.4484
9	.8	.1	.5	1.4	.423	.407	.407	1.237	.4329
10	.5	.7	.4	1.6	.389	.455	.447	1.291	.4435
11	.8	.5	.3	1.6	.398	.439	.423	1.260	.4332
12	.8	.4	.5	1.7	.423	.447	.447	1.317	.4485
1 pm	1.0	.9	.7	2.6	.491	.498	.498	1.487	.4720
2	.6	.4	.4	1.4	.470	.462	.477	1.409	.4902
T_r	20.5	13.2	13.3	47.0	11.841	11.160	11.412	34.413	.4779

APPENDIX TABLE 7. Deviations ($y - \hat{y} = d$) of the observed temperatures y in Table 2 from their expectations $\hat{y}$ as computed from two-term Fourier curves fitted separately to the temperatures for each year.

Year	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Total
43	− .495	− .117	.355	.543	.359	−2.949	2.962	.050	−1.138	−1.309	2.808	−1.067	.002
44	2.261	.152	−1.482	−1.604	2.582	1.218	−2.461	−1.652	3.132	1.504	−4.182	.532	.000
45	1.672	−1.930	.965	−1.757	3.318	−2.918	1.928	−3.003	4.185	−2.277	− .618	.435	.000
46	1.510	−1.613	.692	− .009	.649	−2.096	2.773	−1.887	.191	.959	− .815	− .354	.000
47	1.648	.331	−2.397	1.595	− .253	.789	−1.015	− .781	1.730	.005	−1.163	− .489	.000
48	− .910	.252	.128	−1.157	2.930	−3.534	1.894	.248	− .995	.907	−1.197	1.434	.000
49	1.855	− .326	−2.320	3.637	−1.783	−1.829	3.629	−2.107	− .546	1.480	− .283	−1.404	.003
50	.732	.851	−2.430	1.463	1.163	−2.197	.918	.432	− .620	.421	− .113	− .620	.000
51	− .327	− .563	.103	1.668	−2.203	.414	1.110	− .121	−1.869	2.448	−1.813	1.152	−.001
52	1.226	− .682	.667	−1.316	1.599	−1.059	.290	.282	− .700	.666	.168	−1.141	.000
53	1.104	−1.238	.251	− .531	.381	2.407	−4.937	3.638	− .884	.831	−2.114	1.093	.001
54	1.055	− .101	−1.562	2.180	−1.018	− .236	− .005	.918	− .938	.137	.568	− .997	.001
55	.881	− .310	−2.797	2.779	1.856	−4.511	1.436	2.210	−1.886	.471	− .889	.761	.001
56	−1.202	1.164	−1.191	− .285	.512	2.584	−4.998	2.803	.691	− .448	−2.012	2.383	.001
Total	11.010	−4.130	−11.018	7.206	10.092	−13.917	3.524	1.030	.353	5.795	−11.655	1.718	.008

Payne & Lane Printers, New Haven, 2M-7-58